American Psychological Association

Volume 109 Number 3 April 2017

Published eight times

ISSN 0022-0663

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Steve Graham, at steve.graham@asu.edu. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Journal of Educational Psychology® (ISSN 0022-0663) is published eight times (January, February, April, May, July, August, October, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2017 rates follow: *Nonmember Individual:* $250 Domestic, $292 Foreign, $314 Air Mail. *Institutional:* $953 Domestic, $1,030 Foreign, $1,054 Air Mail. *APA Member:* $123. *APA Student Affiliate:* $75. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology,* American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of

Educational Psychology®

www.apa.org/pubs/journals/edu

April 2017

Volume 109
Number 3

Science & Social Studies

Mathematics

Motivation

Achievement

(*Contents continue*)

Other

Journal of Educational Psychology
2017, Vol. 109, No. 3, 301–320

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000128

Acquiring Science and Social Studies Knowledge in Kindergarten Through Fourth Grade: Conceptualization, Design, Implementation, and Efficacy Testing of Content-Area Literacy Instruction (CALI)

Carol McDonald Connor
University of California, Irvine

Jennifer Dombek, Elizabeth C. Crowe, and Mercedes Spencer
Florida State University

Elizabeth L. Tighe, Sean Coffinger, Elham Zargar, and Taffeta Wood
Arizona State University

Yaacov Petscher
Florida State University

With national focus on reading and math achievement, science and social studies have received less instructional time. Yet, accumulating evidence suggests that content knowledge is an important predictor of proficient reading. Starting with a design study, we developed content-area literacy instruction (CALI) as an individualized (or personalized) instructional program for kindergarteners through 4th graders to build science and social studies knowledge. We developed CALI to be implemented in general education classrooms, over multiple iterations (n = 230 students), using principles of design-based implementation research. The aims were to develop CALI as a usable and feasible instructional program that would, potentially, improve science and social studies knowledge, and could be implemented during the literacy block without negatively affecting students' reading gains (i.e., no opportunity cost). We then evaluated the efficacy of CALI in a randomized controlled field trial with 418 students in kindergarten through 4th grade. Results reveal that CALI demonstrates promise as a usable and feasible instructional individualized general education program, and is efficacious in improving social studies (d = 2.2) and science (d = 2.1) knowledge, with some evidence of improving oral and reading comprehension skills (d = .125).

Keywords: elementary education, reading, science, social studies, design studies

Supplemental materials: http://dx.doi.org/10.1037/edu0000128.supp

Instructional time in the classroom is a precious commodity and, in the early elementary grades, priority is given to establishing strong reading and mathematics skills with little time to focus on content areas such as social studies and science (Banilower, Smith, Weiss, Malzahn, & Campbell, 2013; Duke, 2000; Fitchett, Heafner, & Lambert, 2010; Jeong, Gaffney, & Choi, 2010). Our aim in the two studies reported here was to develop an instructional program that could be provided during the dedicated block of time devoted to teaching literacy and to test the efficacy of this program in improving students' content-area knowledge in social studies and science.

Content-area knowledge has been defined as the knowledge of a particular topic (Hirsh, 2006) and as academic knowledge (Snow, 2010). Although frequently used interchangeably with background or general world knowledge, content-area knowledge focuses on particular areas of disciplinary knowledge, for example science and social studies.

Social Studies

Social studies is defined by the National Council for the Social Studies (NCSS, 1992; http://www.socialstudies.org) as an "integrated study of the social sciences and humanities to promote civic competence" (NCSS, 1994). Knowledge of social studies is important for students' ability to understand their inclusion in history (Alleman & Brophy, 2003) as well as to develop spatial knowledge (Macken, 2003), develop empathy and understanding for human

This article was published Online First September 12, 2016.

Carol McDonald Connor, School of Education, University of California, Irvine; Jennifer Dombek, Elizabeth C. Crowe, and Mercedes Spencer, Florida Center for Reading Research, Florida State University; Elizabeth L. Tighe, Sean Coffinger, Elham Zargar, and Taffeta Wood, School of Education, Arizona State University; Yaacov Petscher, Florida Center for Reading Research, Florida State University.

We thank our collaborators, Christopher Lonigan and the other FSU Reading for Understanding (RFU) investigators, the RFU Network team members, Individualizing Student Instruction (ISI) Lab members past and present including Dr. Barry Fishman, and our school partners, teachers, students, and parents. Funding was provided by the U.S. Department of Education, Institute of Education Sciences, RFU Network Grant R305F100027 and, in part, by R305H04013 and R305B070074 and by National Institute of Child Health and Human Development Grants R01HD48539 and P50 HD052120.

Correspondence concerning this article should be addressed to Carol McDonald Connor, School of Education, University of California, Irvine, 3200 Education Building, Irvine, CA 92697. E-mail: connorcm@uci.edu

activities while being sensitive to temporal and situational aspects (Brophy, Alleman, & O'Mahony, 2003), and understand a chronological span of events, in addition to continuity and change (Hoge, 1996). Moreover, social studies has been cited as a key contributor to citizenship (NCSS, 1994). However, social studies is not assessed consistently nationwide and because social studies is interdisciplinary (e.g., encompassing both history and social science), assessing social studies has been problematic (Risinger & Garcia, 1995). Furthermore, social studies appear to have been left out of the agenda of No Child Left Behind and the new Every Student Succeeds Act (http://www.ed.gov/ESSA), at least more so than other content areas (e.g., science). This has resulted in reduced levels of testing efforts and accountability measures (Grant & Salinas, 2008). However, the inclusion of the requirement that students read and understand content-area texts as part of the Common Core Standards (CCS Common Core State Standards Initiative, 2010) suggests that instruction in how to use and understand social studies texts is essential for mastery of these standards.

Science

Scientific literacy was defined by the National Science Education Standards (1996) as the ability to provide descriptions, explanations, and predictions for naturally occurring phenomena, which allows an individual to engage competently in society. More recently, scientific literacy has been described as follows:

> . . . reading in science requires an appreciation of the norms and conventions of the discipline of science, including understanding the nature of evidence used, an attention to precision and detail, and the capacity to make and assess intricate arguments, synthesize complex information, and follow detailed procedures and accounts of events and concepts. Students also need to be able to gain knowledge from elaborate diagrams and data that convey information and illustrate scientific concepts . . . (Next Generation Science Standards, 2013, p. 1)

Scientific literacy is required for a variety of societal functions, including competency in the workforce, good citizenship, and understanding science in the media (see DeBoer, 2000 for a review). Students who are exposed to explicit science-based literacy instruction during first and second grade demonstrate considerable gains in content-area knowledge in addition to growth in other literacy-based skills (Connor et al., 2012; Romance & Vitale, 2001; Williams, Stafford, Lauer, Hall, & Pollini, 2009). Moreover, the science knowledge gap between students living in poverty and their affluent peers begins early, likely before kindergarten (Morgan, Farkas, Hillemeier, & Maczuga, 2016). Plus, students' science-based content-area knowledge is a strong predictor of their future academic success within the content area (Grant & Fisher, 2010). However, relatively little instructional time is spent with informational (i.e., expository) texts during the earlier grades, and as a result, many students are exposed to fairly limited amounts of science instruction throughout the early elementary years (Banilower et al., 2013; Jeong et al., 2010).

Pearson and colleagues (Pearson, Moje, & Greenleaf, 2010) assert that the lack of content-area instruction in science, which likely holds for social studies as well, may be due to a number of different factors. First, science texts are often less engaging and are not as well written as other texts. Second, instructional time is often spent on text-based activities as opposed to content knowledge. Third, both students and teachers tend to struggle with the concepts, vocabulary, and charts presented in scientific texts. These challenges are further exacerbated by the observation that the time allotted to content-area instruction in science has decreased by about 75 min per week in recent years (McMurrer, 2008); students across the nation are receiving less than 30 min per day of science instruction on average (Blank, 2012). These instructional challenges may explain why fully 35% of eight graders and 57% of students eligible for free and reduced lunch scored below basic levels of science knowledge on the National Assessment of Educational Progress science assessment (National Assessment of Educational Progress, 2011).

The Literacy Block

One likely reason that time spent in social studies and science instruction has decreased is the ubiquitous "literacy block." This is the block of dedicated time that is focused on teaching reading. The literacy block lasts anywhere from 1 to 2 hr depending on the grade level and school. The value of literacy blocks was observed during the late 1990s (Wharton-McDonald, Pressley, & Hampston, 1998) and adopted by many schools during Reading First, a federal program designed to improve students' early reading skills (Gamse, Jacob, Horst, Boulay, & Unlu, 2008). In our collaborations with our partner schools in Florida, Pennsylvania, and Arizona, all principals reported that they have a literacy block.

Within our partner school districts in Florida, where this study was conducted, what can be taught during the literacy block was limited to literacy activities as defined by their core reading curriculum (e.g., Open Court, Reading Mastery, Journeys). This typically excluded teaching science and social studies. Thus, one challenge was to work with our practitioner partners to investigate how to include social studies and science instruction during the designated literacy block.

Associations With Reading Comprehension and Rationale for CALI During the Literacy Block

Past and present research suggests that content knowledge plays a fundamental role in students' ability to comprehend text (Anderson, Reynolds, Schallert, & Goetz, 1977; Voss, Fincher-Kiefer, Greene, & Post, 1986). Previous studies have shown that the quantity (Chi, Fletovich, & Glaser, 1981; Chiesi, Spilich, & Voss, 1979; Dochy, Segers, & Buehl, 1999) and quality of students' content-area knowledge predicts their comprehension abilities (Kendeou & van den Broek, 2005). Quantity refers to the amount of knowledge a student possesses about a particular topic. Research shows that students with more background knowledge have better comprehension abilities than do students with less knowledge (Recht & Leslie, 1988). Quality refers to the correctness of students' background knowledge, which has been shown to influence the amount of understanding gained from text (Diakidoy & Kendeou, 2001). Moreover, gains in comprehension abilities have been found for students who have received explicit classroom instruction in content-area knowledge (Rawson & Kintsch, 2002; Rawson & Kintsch, 2004).

There are a number of models that offer potential mechanisms for conceptualizing the links between disciplinary knowledge (i.e.,

social studies and science) and reading comprehension. For example, the lattice model of reading for understanding suggests that academic knowledge is an integral part of the semantic system. It also hypothesizes that there are reciprocal and synergistic effects of linguistic, cognitive, and text-specific processes that interact with instruction (Child Characteristic × Instruction [C × I] interaction) and with each other to support overall learning (Connor et al., 2014).

The model that most influenced the initial development of CALI was the Direct and inferential mediation (DIME) model (Cromley & Azevedo, 2007). In our adapted model (see Figure 1, top), proficient reading comprehension is supported by decoding and word reading, semantic skills, and background/academic knowledge. Academic knowledge (in this case science and social studies knowledge) theoretically improves strategy use, as well as inferencing and making connections, which indirectly improves reading comprehension. Moreover, following the lattice model, academic knowledge should be an integral part of the semantic system and so should be associated with greater vocabulary as well. The arrows in Figure 1 (top) represent areas that instruction might serve to improve (i.e., are malleable). Hence, in our simplified model (Figure 1, bottom), we conjectured that improving science and social studies knowledge through CALI should improve reading comprehension directly as well as indirectly through stronger vocabulary and oral comprehension.

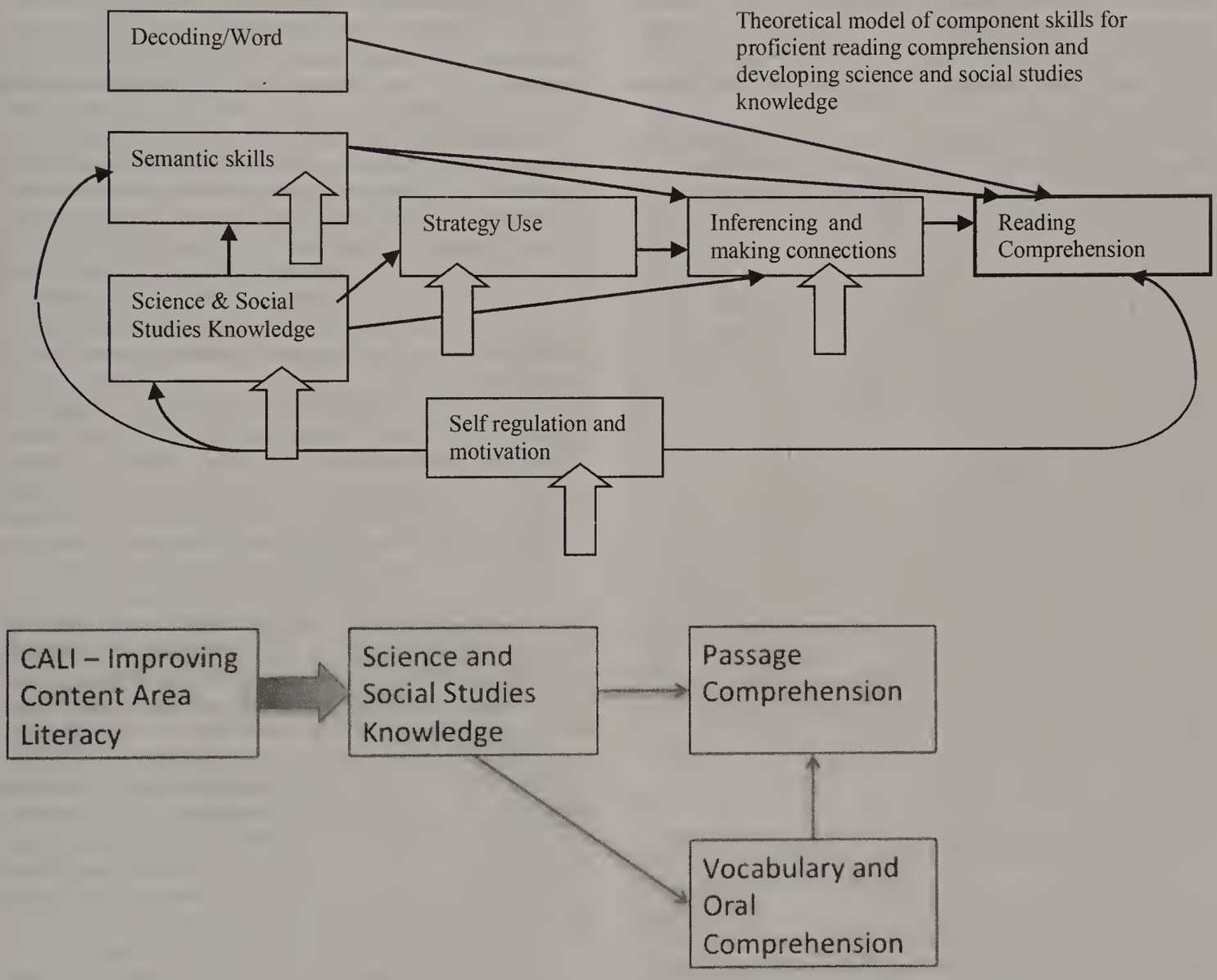


Figure 1. (top) The preliminary logic model. The arrows represent potential content-area literacy instruction (CALI) intervention influences on the components of reading comprehension that should lead to stronger reading comprehension skills (bottom). The simplified version of the theory of change the efficacy study was designed to test. CALI is hypothesized to improve science and social studies knowledge, which in turn is hypothesized to predict oral language and vocabulary, which in turn predicts passage comprehension. Following the Direct and inferential mediation (DIME) model, improved science and social studies knowledge should predict passage comprehension directly. Testing reciprocal effects was beyond the scope of the study. See the online article for the color version of this figure.

Presence of Child Characteristic × Instruction (C × I) Interaction Effects on Content-Area Literacy Learning

There is accumulating evidence that at least some of the variability in students' acquisition of content-area literacy is because the effect of particular types of content instruction depends on children's incoming vocabulary, reading, and background knowledge. In a longitudinal correlational study (Connor et al., 2012), researchers found that science activities where students worked with peers interactively to learn science (e.g., discovery learning) were associated with gains in content-area knowledge. However, this depended on students' previous (i.e., fall) academic and world knowledge. There was a large effect for students with stronger academic content knowledge but diminishing effects for students with weaker knowledge whereby students with the weakest skills showed no knowledge gains at all (i.e., a C × I interaction effect). At the same time, when teachers worked with students interactively, students' demonstrated knowledge gains and there were no C × I interactions.

Such C × I interactions are pervasive in reading and so it is certainly reasonable that we would find C × I interactions for science and social studies content-area learning. Although research examining individual differences in the content areas is just emerging, there is a strong body of evidence that C × I interactions are causally implicated in children's acquisition of word reading and reading comprehension (Al Otaiba et al., 2011; Connor et al., 2013). The child characteristics most strongly implicated in these C × I interactions are word reading skills, reading comprehension, and oral language, particularly vocabulary skills, which are all likely to influence and be influenced by students' academic content knowledge (Snow, Lawrence, & White, 2009). There is some limited correlational evidence of C × I interaction effects in science (Connor et al., 2012) but to the best of our knowledge, none in social studies. Because there is no strong evidence that such C × I interactions are causally implicated in children's learning of content knowledge, testing this is one aim of the studies. At the same time, with the strong evidence in reading comprehension, and the close association between academic content-area knowledge and reading comprehension, we conjecture there are likely to be C × I interactions in social studies and science. To meet these aims, we developed and evaluated CALI as an instructional regime (Cohen, Raudenbush, & Ball, 2003), which we describe more fully below.

Design-Based Implementation Research (DBIR)

Our overall approach in the development of CALI was guided by an educational research framework known as design-based implementation research (DBIR). Building on design-based research (e.g., Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003), DBIR was developed in an effort to better understand the problem of why so many educational interventions are relatively fragile (Fishman, Marx, Blumenfeld, Krajcik, & Soloway, 2004). Design-based research and DBIR are rooted in "small-t theories", as opposed to "capital-T Theories", which may lack details on practical applications. Thus while the capital-T theories framing our studies were the DIME model and C × I interactions, the overarching "small-t theory" guiding our work was examining how the DIME model and C × I interactions, focusing on content knowledge, might operate with younger children in elementary schools.

DBIR aims to create interventions that are aligned with the needs and challenges of schools as organizations, to develop interventions that are more usable and, therefore, likely to be more scalable. There are four key principles behind DBIR. First, there must be a common commitment to solving problems of practice as constructed by educators and educational leaders; that is, from the perspective of those who will ultimately be responsible for implementing interventions. That is why in this project we worked collaboratively on design with educational leaders and teachers who were part of the design team. Second, DBIR engages in iterative, collaborative design of solutions targeting multiple levels of the system: design that is informed by ongoing and systematic inquiry into implementation and outcomes, and is consistent with the research approach in this proposal. Third, there is a common commitment to building theory and knowledge within the research community. And fourth, there is a focus on developing sustainable change within systems.

In conducting DBIR, the procedure involves iterating between design and testing to continually refine the instructional regime toward the aims established by the theories. A number of subtheories may also be at play with respect to individual design elements, guiding designers with respect to how *particular elements* of the intervention function with respect to the aim. We would expect that both our design instances and our "small-t theories" about how those designs function would evolve over time, informed both by usability and outcome data generated through our iterative design pre–posttest studies.

One might argue that what we are describing is design-based research (Anderson & Shattuck, 2012) and we do not disagree. According to Fishman, Penuel, Allen, Cheng, and Sabelli (2013), DBIR builds on design-based research and theories of student learning to "contribute to theories of organizations and institutions . . . by pointing out how the deployment of new tools . . . can bring to light new needs for coordination across different system levels . . ." (p. 144). The literacy block is pervasive throughout school systems with sometimes limiting rules about "what counts" as appropriate literacy instruction. Our aim was that CALI might be a way to test whether we could find ways to expand definitions of "what counts" without interfering with an effective way to teach reading (Wharton-McDonald et al., 1998).

There are not, as yet, many examples of DBIR work. Notable examples include the work of the Strategic Education Research Partnership, an outgrowth of the National Research Council (Donovan, 2013), the development of assessment-to-instruction software (Connor et al., 2011), and earlier work between the University of Michigan and the Detroit Public Schools as part of the National Science Foundation-funded Center for Learning Technologies in Urban Schools (Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000). In addition to developing CALI as an implementable literacy instructional program, we also planned that this project might serve as another example of DBIR in practice, thus advancing the state of the art.

Specific DBIR aims included (a) exploring ways to expand the definition of what is considered acceptable instruction (i.e., social studies and science) to be delivered during the literacy block, (b) investigating ways to build in science and social studies instructional affordances that promote the learning of students with

differing language and literacy skills, and (c) studying ways to feasibly use evidence-based and disciplinary-specific instructional practices from social studies (e.g., original sources) and science (e.g., experiments), and evidence-based literacy instruction practices from literacy research (e.g., assessment-guided instruction, discussion) that would support students' social studies and science knowledge learning from text without opportunity cost (i.e., negatively impacting gains in reading).

The Efficacy Study

Randomized controlled trials (RCTs) are considered the gold standard in education science (Shadish, Cook, & Campbell, 2002; Shavelson & Towne, 2002). This is because alternative explanations cannot be ruled out in pre–post study designs, which are frequently used in design studies and which we used in our DBIR. At the same time, researchers conducting RCTs take a very different perspective than do researchers conducting DBIR and, hence, make different epistemological assumptions; and we acknowledge this tension. For example, while DBIR is essentially contextual, RCTs are based on the assumption that an intervention should be effective across a number of contexts, and when implemented by teachers who did not design the lessons. Indeed, our efficacy trial was conducted in a different partner district than our DBIR studies, with different aims and assumptions, and by teachers who were not on the original design team.

Our justification for this shift in perspectives was that before moving forward in disseminating CALI, we needed to be sure that it was actually effective in improving social studies and science knowledge when delivered during the literacy block in different contexts, by different teachers, and in different schools with different goals and challenges. As an efficacy RCT, in contrast to effectiveness or scale up RCTs, our purpose was to examine whether CALI promoted improved content knowledge when implemented with high fidelity under more controlled but still real-world conditions. Hence, the aims of the efficacy study were fourfold: (a) to evaluate whether CALI would be efficacious in building content-area knowledge in social studies and science; (b) to examine whether the effect of CALI depended on students' incoming language and reading comprehension skills (i.e., C × I interactions); (c) to examine whether CALI could be implemented during the literacy block without opportunity cost—that is, that time spent in content-area literacy instruction would not come at the cost of diminishing language and reading comprehension skills; and (d) to test our logic model (Figure 1, bottom) that improving academic knowledge would improve both oral and reading comprehension.

The Design Study

Method

To pursue our DBIR research and development research aims, we used a mixed-methods approach that combined qualitative and quantitative techniques to inform the iterative development of CALI. We employed observational, interview, and analytic methods during a series of iterations that also employed a pre–post correlational design to help us determine the usability, feasibility and promise of efficacy of CALI and to assess whether there were C × I interactions.

Participants

Design team. The design team included the principal investigator of the project, whose clinical degree was in speech language pathology and whose doctorate degree was in education, specifically special education—language, literacy and culture. At the time of the study, she was an associate professor in developmental psychology, with over 15 years of teaching experience. The project director had a bachelor of science degree in elementary education and a master of science degree in language and literacy, with 4 years of experience working with elementary school teachers and students to build literacy skills. Other members of the team included a student majoring in journalism, a doctoral student in early childhood education, a doctoral student in reading and language arts education, an expert in science education, and three teachers who worked full time on the project as developers and teachers. Additionally, we worked closely with school principals and classroom teachers in our partner schools, gaining their advice and insight through formal and informal interviews. Overall, the team had over 25 years of teaching experience and represented several disciplines.

Participants. The design studies were conducted in 2010–2011, focusing on social studies, and in 2011–2012, focusing on science, in three elementary schools in a North Florida school district. In the 2010–2011 school year, participating students were in kindergarten (n = 35), second (n = 66), third (n = 73), and fourth (n = 56) grades in two schools with over 40% of students participating in the U.S. Free and Reduced Lunch Program (FARL) across 19 classrooms. The students in the sample were highly diverse with 54% African American, 34% White, 8% Hispanic, and 4% belonging to other ethnicities. About half were girls and about 12% qualified for special education services. In the 2011–2012 school year, students in kindergarten through fourth grade in one continuing school and one new school participated. Again, the student sample was diverse and very similar to the students in the previous year. There were 57 kindergarteners, 41 first graders, 51 second graders, 40 third graders, and 38 fourth graders (total n = 227).

Nineteen teachers and three principals in the partner schools participated both formally and informally in the design process. Informal conversations between teachers on the research team and teachers in the schools proved to be most helpful in redesigning CALI. Formal meetings with principals focused on the requirements for including CALI as part of the literacy block.

Assessments

Proximal content knowledge assessments. Used for both the design studies and the RCT, we developed pre–post unit assessments for each unit that included 12 multiple-choice questions and three open-ended questions, and which focused on the topic covered in the unit. During the design studies, the assessments were used to measure how well students were learning the content of the units and whether there were C × I interactions. During the RCTs, the assessments were used as a proximal measure of CALI efficacy.

Two of the 12 multiple choice questions were on topics that were not explicitly covered during the unit. These questions functioned as counterfactual items to provide a more conservative estimate of pre–posttest gains (i.e., we would not expect students to learn this content unless it was covered during regular instruction in the classroom). Reliability on the assessments was acceptable (see Table 1). On the third-grade science Unit A assessment, a typical multiple choice question was as follows (D is correct):

> Which example shows that heat can change an object?
>
> A. Riding a bike down a hill
>
> B. Watching a video in slow motion
>
> C. Drawing a picture of the sun
>
> D. Baking an apple pie

On that same assessment, a typical open-ended question was as follows:

> The sun provides us with several forms of energy. Name a form of energy that the sun provides us and tell how we use it.

The principal purpose of the open-ended questions was to assess how well CALI supported the ability to answer more complex questions and students' ability to talk or write about what they had learned. For all of the open-ended questions (i.e., Items 13–15), two researcher assistants scored each question on a 0–3 point scale, with 3 representing a complete answer. Overall, interrater agreement was excellent (see Table 1).

Because spelling and grammar can influence judgments of writing quality (Graham, Harris, & Hebert, 2011), we carefully considered whether spelling and grammar errors might bias the scoring of the students' written responses to open-ended questions. Kindergarteners and first graders dictated their responses; however, second through fourth graders wrote their responses. A careful review of the students' written answers suggested that very few of them used complete sentences; nor were students requested to use complete sentences. Many responses were lists and short phrases. Hence, we did not consider grammar in the analyses. However, many children provided "interesting" spellings, particularly of science and social studies vocabulary. While we did not count against misspellings, it is possible that the research assistants scoring the open-ended responses could not decode the misspelling or might have given a lower score because of misspellings. Therefore, we counted the number of misspelled words and divided by the number of words in the response to compute the proportion of words misspelled. Reliability (kappa) for identifying misspellings was excellent ranging from .86 to .92 across the three raters. Reliability (kappa) for counting number of words was even higher (.96–.99). When we examined correlations between scores on unit posttests and misspellings, with one exception, none of the correlations were significantly different than zero. There was a weak negative correlation for the science Unit A posttest and spelling on that assessment, $r = -.165$, $p = .012$.

Table 1
Researcher-Developed Test Reliability Across All Grades and Conditions

Assessment	α	Inter-rater correlation/κ		
		Item 13	Item 14	Item 15
Social studies Unit A	.85	.93/.87	.98/.97	.95/.94
Social studies Unit B	.79	.88/.79	.96/.88	.90/.86
Science Unit A	.85	.99/.98	1.0	1.0
Science Unit B	.80	.96/.99	.97/.98	.97/.96
Reading-2-Comprehension	.85			

Note. Interrater correlations significant at $p < .001$.

Standardized measures. We also administered the Picture Vocabulary, Letter-Word Identification, and Passage Comprehension subtests of the Woodcock-Johnson-III Tests of Achievement (WJ-III, Woodcock, McGrew, & Mather, 2001). These assessments were administered prior to beginning of the design study to examine C × I interactions. On average, students were achieving grade-level expectations with standard scores averaging 99.5 ($SD = 15$) on Picture Vocabulary and 98 ($SD = 10.7$) on Passage Comprehension. All three measures are psychometrically strong with reliabilities (alpha) between .70 and .98 according to the technical manual.

Procedures

Already developed design components. As described in Connor et al. (2010), a second-grade CALI science unit had already been developed using the 5-E learning cycle (that is, Engagement, Exploration, Explanation, Elaboration, Evaluation; Bybee, 1997) and was used as the foundation for the development of CALI social studies and science across grades. In this version of CALI, each student had a scientist notebook, which was a loose-leaf binder. Modeled after the work of Palincscar and Magnusson (2001) and *Seeds of Science, Roots of Reading* (Lawrence Hall of Science, 2007), students kept all written work, including graphic organizers and responses to questions, in these binders and referred to them throughout the lessons. We also used hands-on experiments and trade science texts including books from *Seeds of Science, Roots of Reading.* Text accommodations had to be made for students with weaker reading skills. Students were grouped according to oral reading fluency and passage comprehension scores and sat with their group during the lessons. The teacher began each lesson with a discussion with the entire class, which frequently included an initial reading of the texts and a review of the work for the day. Then, working in their groups, students would read the book (each student had their own book) or conduct their experiment and would complete their worksheets (e.g., responding to questions; graphic organizer; recording observations). The teacher would float among the groups providing extra support where needed.

Iterative design procedures. The design team met weekly and in these work circle meetings, we reviewed the literature, made initial design decisions and developed lessons, reviewed lesson plans and materials as they were developed, and as CALI was implemented, reviewed conversations with school teachers and principals and the results of the pre–posttests. All of this information, recorded using field notes, summary documents, and documented minutes during the work-circle meetings, was used to make decisions about how to improve the lessons and materials (see decision rules below). Usability (e.g., lesson plan and mate-

rials ease of use) and feasibility (e.g., able to implement in context without undue burden) were both considered.

Initial design decisions. We started with an overall design framework based on our previous research (Connor et al., 2010) and the expertise of the design team. Reviewing minutes from the meeting and summary documents, we made the following decisions in September of 2010. Later relevant changes from other documents are added as "Notes" below.

1. CALI will be implemented in 2- to 3-week units, 4 days/week, with each lesson lasting 1 day.

2. Units will use a common framework that builds on previous research. We decided that we would have to adapt this framework for social studies and, if it was effective, then translate it to use with science. Therefore, we initially developed a five-phase system: *connect, clarify, research, apply, and appraise,* which was modified to a four-phase system based on findings (see Figure 2 for final framework).

    a. In the *connect* lessons, students will connect a concept in social studies (e.g., state government) with something that is current, in their life, or in the news (e.g., the current governor). The idea is to begin to build the concept while building enthusiasm and motivation.

    b. *Clarify* lessons will focus on reading and how to read and learn from secondary sources in social studies. These lessons tie back to the *connect* lesson to maintain enthusiasm and motivation, and help students continue to feel connected to the topic.

    c. The *research* lessons will teach children about primary sources (photographs, journals, letters) and how to read and use them to elaborate on secondary sources (textbooks). For science, this would be experiments.

    d. The *apply* lessons will focus on making connections and drawing conclusions through projects (e.g., posters) and writing. The goal is that children will learn the concepts covered in each unit as well as how to read and learn from expository text.

    e. In the *appraise* lessons, teachers and students will reflect on what they had learned (notes from October 11, 2010).

        i. Note. This phase was ultimately dropped and incorporated into the *apply* lessons.

3. Teachers will use specific discussion strategies, such as brainstorming and think-pair-share, to promote students' engagement and learning.

4. Instruction will be semiscripted for teachers (sample lesson plans are provided in the online Supplementary Materials).

5. Lessons and materials will be individualized for flexible learning groups. We planned to test whether grouping students on reading comprehension or some other skill, such as pretest, would be most effective. Based on our previous research, reading comprehension was the most likely candidate and we started there.

6. Topics were selected based on Common Core State Standards, which had just been published in draft form, and the Sunshine State Standards in Florida.

    a. Note. We also consulted school principals and the deputy superintendent in charge of elementary curriculum. The clear message was that if the topics did not align with the state standards and the new core standards, CALI would not be usable or feasible. The topics selected are provided in Table 2. The educational leaders also made it clear that *only* if the principal goals of CALI were reading and understanding expository text, would CALI be acceptable for instruction during the literacy block.

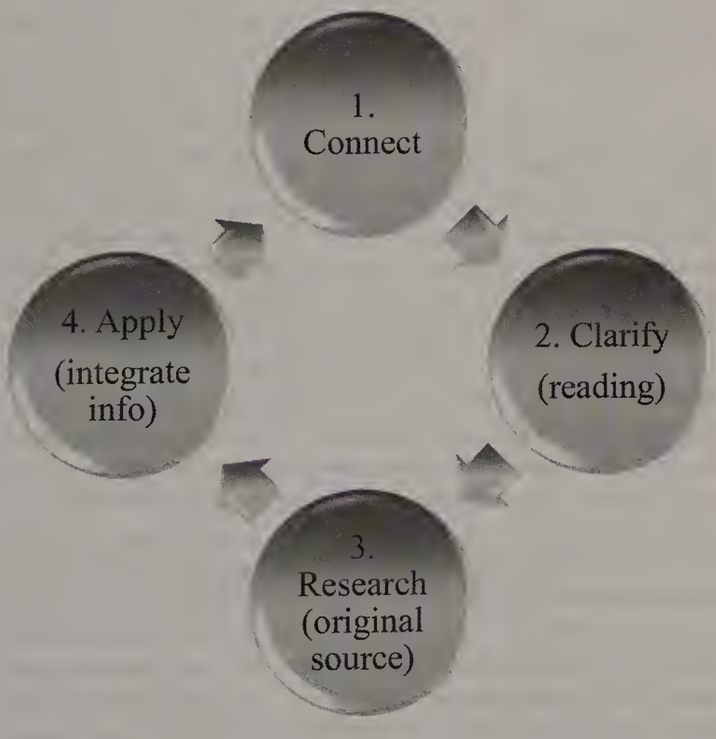


Figure 2. Organizational framework for content-area literacy instruction.

That document also detailed our plan regarding how we would make decisions to change or leave the intervention as is.

a. Most important—gains on key indicators will indicate that we did not need to change the intervention substantially.
b. Usability and feasibility—reasonably easy to for teachers to implement and follow scripts; ability to complete lessons in allotted time, with students engaged, would suggest that intervention is about "ready for prime time."
c. Grouping—interactions among students are appropriate; no C × I interactions on posttests.

Finally, the document provided a logic model (see Figure 1, top). Again, this model was adapted from the DIME model (Cromley & Azevedo, 2007) and the proposed component model of

Table 2
Topics for Content-Area Literacy Instruction Social Studies and Science

Grade	Subject	Unit	Topic
Kindergarten	Social studies	A	Rules and laws
			George Washington
		B	Community helpers
			Transportation
	Science	A	Food groups
			Food groups
		B	Five senses
			Observation, Using the five senses
1	Social studies	A	Economics: Needs and wants
			Economics: Goods and services
		B	Eleanor Roosevelt
			Rules, laws and being responsible members of a community
	Science	A	Living and nonliving things
			Plants: Plant parts and plant growth
		B	Earth's surface
			Earth and space
2	Social studies	A	Abraham Lincoln
			Constitution: Laws and voting
		B	Economy: Goods, services, income and importing
			Economy consumer and producer relationship
	Science	A	Scientific method
			Observing patterns in nature: weather
		B	Matter
			States of matter
3	Social studies	A	Citizens: Rights and responsibilities
			Thomas Jefferson
		B	Environments and communities
			Natural resources and conservation
	Science	A	Stars
			Energy
		B	Plants
			Animals
4	Social studies	A	Branches of U.S. government
			The U.S. Constitution
		B	Economy: Goods, services, consumers and producers
			Transcontinental railroad
	Science	A	Earth in space and time
			Rocks
		B	Interdependence
			Earth's natural resources

language (Connor et al., 2014), and informed our initial version of CALI. The theory of change model proposes that science and social studies instruction supports developing semantic and academic knowledge, whereas the ability to make strong inferences and connections within and between texts supports reading comprehension. In turn, stronger reading comprehension of expository texts supports the development of science and social studies knowledge and the semantic system overall. Together, the model conjectures a reciprocal loop that supports proficient reading for understanding. In this model, decoding is an important source of influence on students' comprehension and so some focus on decoding words (e.g., multisyllabic and compound words found in science and social studies texts) would be provided to students with weaker skills. However, decoding was not a principal focus of CALI.

Schedule of implementation. We focused on kindergarten and second through fourth grade and social studies for the first iteration of the CALI. First graders in our partner schools were already involved in another study and so designing lessons for them was postponed until the following school year. Preintervention assessments (the WJ-III assessments and the Unit 1 and 2 tests) were conducted the first week of November. Dividing the students into two cohorts, we implemented Unit 1 in half of the classrooms during the second week of November. We then reviewed how lessons were implemented during work circle meetings, made necessary revisions, and implemented the same unit to the second cohort of students during the third week of November. We then reviewed the implementation again and made any needed changes. We repeated this procedure for Unit 2 through the end of November and the beginning of December. Following the implementation of Unit 2, we conducted the postintervention assessments (Unit 1 and 2 Posttests). Having two cohorts allowed us flexibility and the ability to make and test changes to the lesson plans and materials more quickly. We then reviewed the results of the first set of studies following our decisions rules. One of the first decisions we made was to conduct pre- and posttesting for each unit. In this way, through the 2010–2011 and 2011–2012 school years, we developed CALI social studies and science.

Design Study Results

Iterations. Data results for the DBIR study are provided in Table 3. During our *first iteration*, there was a large effect of treatment on pre–posttest gains on the unit content assessments, taking into account the counterfactual items (d = .84), however, we found C × I interaction effects. Specifically, students who started the intervention with weaker pretest unit scores made greater gains pre- to posttest, which we decided was acceptable. We also found that students who had stronger passage comprehension scores made greater gains than did students with weaker scores, which was not acceptable. When the team teachers implemented CALI, they reported that the unit required 3–4 weeks to accomplish and that 2 weeks was too short. However, given school principals' and teachers' feedback about their calendars, 4 weeks would be too long and would be interrupted by school holidays, assessment weeks, and other school activities.

Based on these results and feedback from the teachers who implemented CALI, the team (including the teachers) made the following decisions: (a) we decided to design 3-week units; (b) to revise the lessons to include more writing, particularly in response to open-ended questions; (c) to change the format moving from primarily whole-class implementation with the teachers floating among groups to starting each lesson with a whole class lesson and then moving to small flexible learning groups that rotated through a teacher table; additionally, (d) the research teachers stated that the commercially available leveled texts (by Pearson Scott Foresman) we were using were inadequate inasmuch as the lower level (i.e., easier) texts were not including key vocabulary and information, which, they conjectured, might contribute to the C × I interactions. When the entire team (including the teachers) reviewed the texts, they concurred. Based on this observation and after further discussion (and encouragement from our journalism student), the design team decided to write leveled readers rather than use the trade books. For each unit we wrote one or two books for each group—blue (above grade level), green (below grade level) and yellow (at grade level, see online Supplementary Materials). We relied on metrics from Lexiles (https://lexile.com/) and Coh-Metrix (http://www.cohmetrix.com/) along with professional judgment and trial and error to develop the text. Each student received the books for the unit and kept them in the scientist notebook.

In the second iteration, we tried grouping students by unit pretest score. Review of the data revealed that, again, students with weaker preunit test scores made greater gains. Fortunately, we found no C × I interaction with passage comprehension; however, we found that students with stronger vocabulary scores made greater gains than did students with weaker vocabulary scores. Discussion during the work circle meetings then focused on how we might incorporate explicit vocabulary instruction into the lessons themselves. The researchers and research teachers agreed that we might highlight key vocabulary in the texts themselves and add specific vocabulary discussions to the scripts. In response, the team revised and expanded use of leveled books and highlighted key vocabulary, expanded the use of graphic organizers, and developed explicit vocabulary instruction using the key vocabulary words that were incorporated into the lessons.

Finally, the teachers who implemented CALI reported that using the unit pretest left the groups too heterogeneous with regard to reading skill, and the leveled materials (e.g., texts, graphic organizers) were not appropriate for all members of the small group. In their opinion, using the passage comprehension score to create flexible learning groups resulted in more feasible implementation of individualized instruction.

During the second iteration, research teachers experimented with a number of different grouping strategies by varying the size, number, and skill to determine group membership. They reported that the optimal group size appeared to be three to five students and that the groups were too large with six students. There was discussion during the work circles that this might mean there would be uneven numbers of higher performing, typical, and lower performing student groups. The team decided that this was acceptable and we would try out smaller groups during the third iteration.

In our *third iteration*, we used groups that were created based on passage comprehension scores and that had no more than five students per group. Work-circle discussions during the third iteration focused on improving usability and feasibility. Our team teachers who were implementing CALI suggested several ideas for improving the usability of the scripts including discussion questions and strategies. Specifically, they suggested that we highlight the strategies for improving student participation that they found most useful—think, pair share; brainstorming; questioning; and so forth—and include definitions on the lesson plans. Additionally,

Table 3
Design Study Results Across Four Iterations of Social Studies

Iteration	Units pretest (z score)	Units posttest (z score)	Units pretest counterfactual (CF; z)	Units posttest CF (z score)	Effect size (d) gains; target vs. CF
Iteration 1	−.47	.47, SD = .90	−.16	.16	.84
Iteration 2	−.48	.49, SD = .87	−.05	.05	.87
Iterations 3 and 4					.86

Testing for Child × Intervention interactions	Pretest correlated with gains (r)	WJ Passage Comprehension SS correlated with gains (r)	WJ Picture Vocabulary SS correlated with gains (r)
Iteration 1	−.425***	.345*	*ns*
Iteration 2	−.433***	*ns*	.298*
Iterations 3 and 4	*ns*	*ns*	*ns*

Note. WJ = Woodcock-Johnson; SS = standard scores; *ns* = not significantly different from 0.
* $p < .05$. *** $p < .001$.

the team realized that looking ahead to the RCT, we needed to develop a way to provide CALI without relying on the initiating whole class discussions because children were going to be randomly assigned within classrooms. Additionally, having an alternative to the whole class approach might better support the new district mandates (per the partner principals) that at least 50% of the literacy block had to be conducted in small groups. Thus, the team worked together to create a small flexible learning group rotation so that teachers met first with the group of students with the weakest reading comprehension scores (i.e., the green groups) and then met with the other groups. The design team also conjectured that this more structured approach might allow the participation of paraprofessionals although this was not tested.

The final version of CALI, where all instruction was provided in small groups, was tested in the *fourth iteration* and was used for the RCT. In the fourth iteration, a review of the data revealed strong pre–post unit test effects ($d = .86$) and there were no significant C × I interactions. Teachers reported that holding discussions during the small groups was feasible. Plus, the format provided opportunities for children who were reticent to talk more opportunities to participate. The length of each group meeting was longer—closer to 15–20 min per group—to cover the entire lesson with each group, compared to the protocol that started with whole-class discussion, which took about 30 min for the entire class.

After four iterations, which took the entire 2010–2011 school year, we developed two social studies units (see Table 2 for topic and online Supplemental Materials for sample lesson plans; materials available upon request from the first author). Development of the CALI science units and first grade social studies occurred during the first half of the 2011–2012 school year and we modified our DBIR approach so that we implemented CALI science Unit A and then developed CALI Unit B based on what we learned from implementing CALI Unit A, and then retroactively revised CALI Unit A based on what we learned from implementing Unit B. Our rationale was that we had learned important information in the four iterations that we conducted for CALI social studies that could be applied to CALI science. With this accelerated development, CALI science was developed in time to be in the efficacy trial once CALI social studies was completed.

Data results from CALI science suggested more moderate effects of treatment for Unit A ($d = .58$) with a C × I interaction with pretest, which was negative ($r = -.29$, $p < .01$, i.e., students with weaker pretest scores made greater gains) but a positive correlation with passage comprehension, $r = .29$, $p < .001$ suggesting that students with stronger passage comprehension scores made greater gains on the unit assessment. We made changes to the lesson plans and materials to address the C × I interaction with passage comprehension. The final iteration indicated that CALI showed evidence of both promise and feasibility without significant passage comprehension C × I interaction effects.

Description of CALI Protocols, Lessons, and Materials

Protocol. Each 3-week unit was provided during the literacy block, 4 days per week for 30 min when initiating lessons with the entire class and between 15 and 20 min per group when using small groups. The length of the components varied but, in general, *connect* lessons were 1 day, *clarify* lessons were 3–4 contiguous days, *research* lessons were also 3–4 contiguous days, and *Apply* lessons were 3–4 contiguous days. CALI science was developed as the flexible learning group version but could easily be adapted to the whole class protocol. Selected lesson plans and materials are provided in the online Supplemental Materials.

Flexible learning groups. The team reached consensus that assigning by reading comprehension skill level across three groups—below grade expectations (green, standard score approximately below 90), about at grade level (yellow), and above grade level (blue, standard score approximately above 110) was feasible and supported students' content literacy learning without C × I interactions that negatively impacted students with weaker incoming skills. To keep the number of student per group at the optimal numbers of no more than five students, some classrooms had more than one group at each level (e.g., two green groups, two yellow groups, and one blue group).

Leveled books, graphic organizers, and scientist notebooks. The leveled short books written for the units appeared to be critical for reducing C × I interactions (see results for the third and fourth iterations). Again, the team found that leveled trade books tended to delete important content and vocabulary to make the books easier to read, which we assumed to be the reason we found C × I interactions effects for vocabulary in the second iteration. Additionally the books could become part of the scientist's notebook (Palincsar & Magnusson, 2001) so the students could access them easily throughout the lessons. We also developed leveled graphic organizers for each group; again making sure that content and vocabulary were the same across groups but with more built-in scaffolding for weaker readers. For example, students in the green groups might write three sentences in a graphic organizer whereas students in the blue group would be expected to write five sentences. Scientist notebooks were loose-leaf binders and students kept their books, graphic organizers, and notes in these books. Observations revealed that students used their notebooks during lessons and referred to previous work to make inferences about new content.

Use of original sources in social studies and experiments in science. A key part of CALI was helping students read across various types of texts and to learn disciplinary-specific practices. Hence, in addition to the leveled books, students read original sources in social studies, such as facsimiles of letters and documents (e.g., Bill of Rights), as well as examined photographs of the time being studied during the *research* lessons. With regard to science, lessons explicitly taught the scientific method, including observation and experimentation. During the *research* lessons in science, students conducted experiments and analyzed data (e.g., graphing observations). Based on our observations, team teacher reports, and pre–post assessments, these strategies appeared to function as anticipated.

Scripting and supporting discussion. Based on team and partner teacher feedback, we developed open-ended scripting that provided specific suggestions for fostering discussion (e.g., open-ended question prompts) but not so much scripting that implementation became cumbersome. The final scripting used in the lessons was, in the design teams' opinion, sufficient to give teachers a good idea of how to implement the lessons while still allowing them professional discretion to elaborate or build on the ideas in the lesson.

CALI Efficacy Study

The DBIR study, although critical to developing CALI, could not demonstrate that CALI was efficacious in other contexts. Indeed, it is possible that the observed gains were context/school specific or the result of our partner classroom teachers' effective instruction, and not due to CALI. Moreover, without a control group, we could not examine whether content knowledge gains differed for students who did or did not participate in CALI. Nor could we evaluate whether there might be an opportunity cost to implementing CALI during the literacy block, which would manifest as weaker gains in reading for students in the CALI group. Williams and colleagues (2009) note that one reason social studies and science are taught less is because there is a perception that time spent in social studies and science is less time spent in reading instruction. Therefore, our next step before scaling up with an effectiveness RCT was to conduct an RCT efficacy study.

Method

Participants. Kindergartners through fourth graders (n = 418) in 40 classrooms attended six schools in a large school district located in the Panhandle of Florida. Schoolwide percent of FARL ranged from 40% to 91% with a mean of 57%. After classroom rosters were obtained, students were matched on *passage comprehension* (described below) and randomly assigned within classrooms to participate in CALI or to a business as usual control. Students were then assigned to groups: green (passage comprehension below grade level, standard score [SS] < 90), yellow (at grade level, SS between 90 and 110) or blue (above grade level, SS > 110) groups and remained in these groups throughout the school year.

Of the original 459 students (n = 41), 8.9% left the study before the end of the school year because they moved out of the school or district. The final sample size of 418 students provided a minimally detectable effect size of 0.25 for the overall sample and 0.35 by grade (power = .82, groups = 2 [or 8 by grade], G-Power version 3.1.9.2 and Optimal Design version 3.01). No parents withdrew their child from the study. Attrition was evenly distributed across the treatment and control conditions, $\chi^2(1) = .047, p = .878$, and grade $\chi^2(4) = 1.35$, $p = .878$. We also examined whether there might have been differential attrition based on fall reading comprehension scores and found none, $F(1, 455) = .554$, $p = .457$, for condition by missing-by-spring interaction effect.

In the final sample, there were 212 in the CALI condition and 206 in the control condition and between five and 22 participating students per classroom with a mean of 10 students/classroom. There were 83 kindergarteners, 109 first graders, 75 second graders, 76 third graders, and 75 fourth graders. Twenty-five percent were assigned to the higher performing blue groups, 20% to lower performing green groups, and 55% to the average performing yellow groups. Seventy-seven percent were white, 10% were African American, and the rest belonged to other ethnicities. About 50% of children qualified for FARL, a widely used indicator of family poverty. The distributions were similar for both conditions.

Measures. We used the same measures as described in the design studies with some changes. We added the WJ-III Oral Comprehension test (Woodcock et al., 2001); we did not administer the Letter-Word Identification test, and we added a researcher-developed measure of reading comprehension, Reading-2-Comprehension (R2C). Reliability and interrater reliability for the researcher-developed assessments are provided in Table 1.

The Oral Comprehension test, which was administered live, is designed to measure students' ability to listen to and comprehend a passage and then supply a missing word using syntactic and semantic cues. According to the WJ-III technical manual, split-half reliability for Oral Comprehension is .85 (McGrew & Woodcock, 2001).

We developed a comprehension assessment designed to require inferencing and comprehension monitoring, which was administered to third and fourth graders. Called the Garden Path Maze when developed, and currently named R2C, students are presented a paragraph that has a missing word at the beginning of the paragraph with four possible answers. All four words provided are correct in the context of the first few sentences. Students must finish reading the passage to determine the correct word to select (see the online Supplemental Materials for an example). Rasch analyses suggest good reliability (α = .853 with this sample) particularly for students with scores between −.5 and + 1.8 logits (see Figure S.1 in the online Supplemental Materials for item characteristic curves).

CALI condition. Because this was an efficacy trial, which is conducted in schools but under more controlled condition than an effectiveness trial, CALI was implemented by teachers who were hired by the research team rather than the classroom teacher. All teachers had experience working with children and were either certified or in university programs to become certified. Design team teachers and researchers who had been involved in the development of CALI, as well as a postdoctoral fellow in education with over 7 years of teaching experience conducted professional development for the teachers. The teachers attended a full-day workshop where the aims of CALI and how to implement the lessons were carefully described. Teachers also received hands-on opportunities to implement CALI. The teachers also participated in weekly project meetings and provided feedback and observations regarding CALI implementation.

Fidelity. To determine whether CALI was being implemented as intended, we conducted fidelity observations during implementation. The project director observed each teacher during the first week of implementation and completed a lesson plan checklist. This was to ensure that teachers were generally following the scripts, working with each group; that each group was no more than five children; and that discussion strategies, leveled readers, and graphic organizers were being used as intended. If elements of the lesson were missing or not implemented optimally, the project director provided feedback and observed the teachers again the following week until they were implementing CALI as intended. Then they were observed monthly. No teacher required more than two observations during the first weeks of implementation to achieve adequate fidelity. Based on observations during the first weeks and monthly observations thereafter, the teachers implemented CALI as intended using the small learning group (i.e., groups of no more than five students), during the literacy block in the classroom or in a quiet place near the classroom. Classroom teachers at the participating schools were required to provide small group instruction during the literacy block. Thus, based on team teachers' and classroom teachers' report, CALI was usable, feasible, and well-suited to the organization of the classrooms.

Control condition. Students who were not randomly assigned to participate in CALI received business as usual instruction during the literacy block from the classroom teacher. Based on observations conducted by the researchers, this was principally the scope and sequence of the literacy core curriculum, Houghton Mifflin, which varied depending on the grade. Although some expository text use was observed, there was no focused science or social studies instruction during the literacy block. In general, the quality of instruction was adequate to excellent based on observation. This was expected given the focus in the district on professional development and providing evidence-based reading instruction.

Of note, classroom teachers reported that they taught science and social studies following the Florida Sunshine State Standards but not during the literacy block. Unfortunately, it was beyond the funding available to observe science and social studies instruction although we reviewed the science and social studies cores used (Scott Foresman). Hence, for purposes of RCT fidelity, we might assume that all students, treatment and control, received the content covered in the core curriculum, which would cover the content in CALI because CALI was carefully aligned with the Florida Sunshine State Standards. However, students in CALI would have received more time in science and social studies instruction (CALI plus the cores) and, arguably, less time in core reading instruction than the students in the control group.

Results

Establishing baseline equivalency. Descriptive statistics of pre- and poststudy assessments are provided in Table 4. Analyses of group differences on pre-CALI measures confirmed that there were no significant differences between CALI (i.e., treatment) and control groups at baseline and that, on average, students were performing at expected levels.

Treatment effects on content-area knowledge. We used hierarchical linear modeling (HLM) analyses, to account for the nested structure of the data (children nested in classrooms), controlling for pretest unit scores and grade level. We found significant treatment effects of CALI for both social studies ($g = 2.27$) and science ($g = 2.10$). Students randomly assigned to CALI achieved significantly higher social studies and science postunit scores on the proximal measures. Model results are provided in Table 5.

Because students were given CALI content knowledge assessments prior to and after each 3-week unit, we were able to examine changes and accumulation in social studies and science performance over time across the school year by conducting longitudinal piecewise HLM analyses (Raudenbush & Bryk, 2002) with repeated measure over time nested in students nested in classrooms. We were able to account for the shared within-child variance on the assessments, which was important because of potential form (all assessments followed the same format) and practice effects, and we were able to account for within-classroom shared variance.

Keeping in mind that these assessments were given orally to kindergarteners and first graders whereas second through fourth graders were expected to read and write, we coded kindergarten to first grade (K–1st) = 0 and second to fourth grades (2nd–4th) = 1 and entered the variables at the classroom level (see Table 6). As can be seen in Figure 3, on average, students in the CALI condition showed significantly greater gains with resultant higher scores on the social studies posttest compared to the control ($d_{\text{K–1st}} = .66$; $d_{\text{2nd–4th}} = 1.07$). They also showed higher scores on the science pretest suggesting some transfer across content areas ($d_{\text{K–1st}} = .53$; $d_{\text{2nd–4th}} = .31$). Finally, they made significantly greater gains with large effects of CALI on the science posttests ($d_{\text{K–1st}} = .94$; $d_{\text{2nd–4th}} = 1.17$).

We calculated the misspelling ratios (number of words misspelled divided by the total number of words) for each unit posttest and entered the ratio into the model at the second level (student level). On average, the ratio of misspellings ranged from .06 for the posttest of science Unit B (the last unit implemented) to a high of .13 on the posttest for social studies Unit A (the first unit implemented). In a

Table 4
Preintervention (Pre, Top) and Postintervention (Post, Bottom) Sample Sizes and Means for Pre- and Postintervention Assessments by Content-Area Literacy Instruction (CALI; Treatment) and Control Groups

Measures	CALI		Control	
	N	*M* (SD)	*N*	*M* (SD)
Baseline				
Oral Comprehension (SS)	232	104.9 (12.1)	227	104.3 (12.1)
Vocabulary (SS)	232	100.4 (10.4)	227	101.3 (10.1)
Passage Comprehension (SS)	232	100.5 (12.8)	227	98.8 (12.4)
Preunit social studies knowledge (RS out of 25)	228	11.1 (4.8)	219	11.1 (4.7)
Preunit science knowledge (RS out of 32)	214	15.0 (6.0)	218	14.3 (6.0)
Postintervention				
Oral Comprehension (SS)	212	106.8 (12.9)	206	106.7 (11.8)
Vocabulary (SS)	212	102.4 (9.1)	206	101.9 (9.5)
Passage Comprehension (SS)	212	100.4 (13.4)	206	101.1 (11.6)
Preunit social studies knowledge (RS out of 25)	220	22.7 (5.51)	215	12.9 (5.0)
Preunit science knowledge (RS out of 32)	212	25.06 (5.4)	211	16.5 (6.0)

Note. SS = standard scores; RS = raw scores. Based on 95% confidence intervals of means, there were no significant differences between CALI and control groups prior to implementing the intervention ($p > .05$). SS are reported to facilitate interpretation of scores where the grade-corrected mean is 100 (15). Developmental scores (W) were used in analyses.

Table 5

Hierarchical Linear Modeling Results for Content-Area Literacy Instruction (CALI) Effects on Social Studies and Science Proximal Posttests Controlling for Grade Level and Pretest Scores

	Social studies					Science				
Fixed effect	Coefficient	*SE*	*t* ratio	Approximate *df*	*p* value	Coefficient	*SE*	*t* ratio	Approximate *df*	*p* value
Fitted mean Control posttests, β_{00}	11.737	.513	22.871	38	<.001	15.596	.522	29.838	38	<.001
Grade level, β_{01}	.544	.204	2.669	38	.011	.582	.209	2.785	38	.008
Effect of CALI, β_{10}	9.702	.398	24.319	393	<.001	7.854	.360	21.817	381	<.001
Effect of pretests, β_{20}	.770	.051	14.883	393	<.001	.718	.035	20.188	381	<.001

	Social studies					Science				
Random effect	*SD*	Variance component	*df*	χ^2	*p* value	*SD*	Variance component	*df*	χ^2	*p* value
INTRCPT1, r_0	1.00283	1.00567	38	60.69455	.011	1.37131	1.88049	38	94.75210	<.001
Level 1, *e*	4.14900	17.21418				3.67617	13.51421			

Note. Students in CALI = 1, control = 0; kindergarten = 0, first grade = 1, second = 2, third = 3, and fourth = 4. Pre- and posttests are reported in raw total scores. Deviance for social studies = 2,492.383; deviance for science = 2,338.576. Posttests totals$_{ti}$ = β_{00} + β_{01}* Grade level$_i$ + β_{10}* CALI$_{ti}$ + β_{20}* Pretest totals$_{ti}$ + r_{0i} + e_{ti}.

series of models, only misspelling on the science Unit A posttest (M = .11, SD = .20) predicted the outcome (coefficient = −2.91, $p < .001$). However, when the other misspellings were trimmed and misspelling on science Unit A added to the slopes (both social studies and science), misspelling had no significant effect on students' gains in content knowledge ($p = .117$) and there continued to be a CALI treatment effect (see Table S.3 in online Supplemental Materials).

Testing for Child × Instruction interactions. A key aim of CALI's design was to reduce C × I interaction effects. We tested for C × I interactions of oral comprehension, vocabulary, and passage comprehension in two ways: First by examining interaction effects in our piecewise growth curve model and then using quantile regression (Koenker & Bassett Jr, 1978; Petscher & Logan, 2014). A description of quantile regression and the results

Table 6

Hierarchical Linear Modeling Longitudinal Piecewise Results for Content-Area Literacy Instruction (CALI) Unit Tests Considering Kindergarten and First Grade (K1 = 1) and Second to Fourth Grade, Where Grades 2–4 Represent the Fixed Reference Group (= 0)

Fixed effect[a]	Coefficient	*SE*	*t* ratio	Approximate *df*	*p* value
Mean SocS unit test score at time 0	3.47	.207	16.769	38	<.001
K1 effect	2.60	.417	6.235	38	<.001
CALI effect at Time 0	−.08	.250	−.336	422	.737
CALI × K1 effect at Time 0	−.74	.446	−1.671	422	.096
SCI effect on intercept	1.76	.605	2.904	2,882	.004
K1 effect	.30	1.586	.189	2,882	.850
CALI effect	−5.78	.873	−6.615	2,882	<.001
CALI × K1 effect	3.39	1.298	2.613	2,882	.009
Change per week (i.e., slope for SocS)	.26	.034	7.568	38	<.001
K1 effect	−.008	.044	−.187	38	.853
CALI effect (i.e., treatment effect)	.64	.051	12.500	2,882	<.001
CALI × K1 effect on slope	−.24	.067	−3.593	2,882	<.001
SCI Slope	−.16	.051	−3.109	2,882	.002
K1 effect on SCI slope	−.049	.090	−.537	2,882	.591
CALI effect on SCI slope	−.09	.075	−1.223	2,882	.221
CALI × K1 effect	.022	.110	.203	2,882	.839

Random effect	*SD*	Variance component	*df*	χ^2	*p* value
Level 2[b]	1.67166	2.79446	408	1,697.69422	<.001
Level 1[b]	2.60447	6.78326			
Level 3 intercept[c]	.36609	.13402	38	49.02520	.109
Level 3 slope[c]	.06618	.00438	38	122.55948	<.001

Note. SocS = social studies; SCI = science. Deviance = 17,023.13.
[a] Final estimation of fixed effects (with robust standard errors). [b] Final estimation of Level 1 and Level 2 variance components. [c] Final estimation of Level 3 variance components.

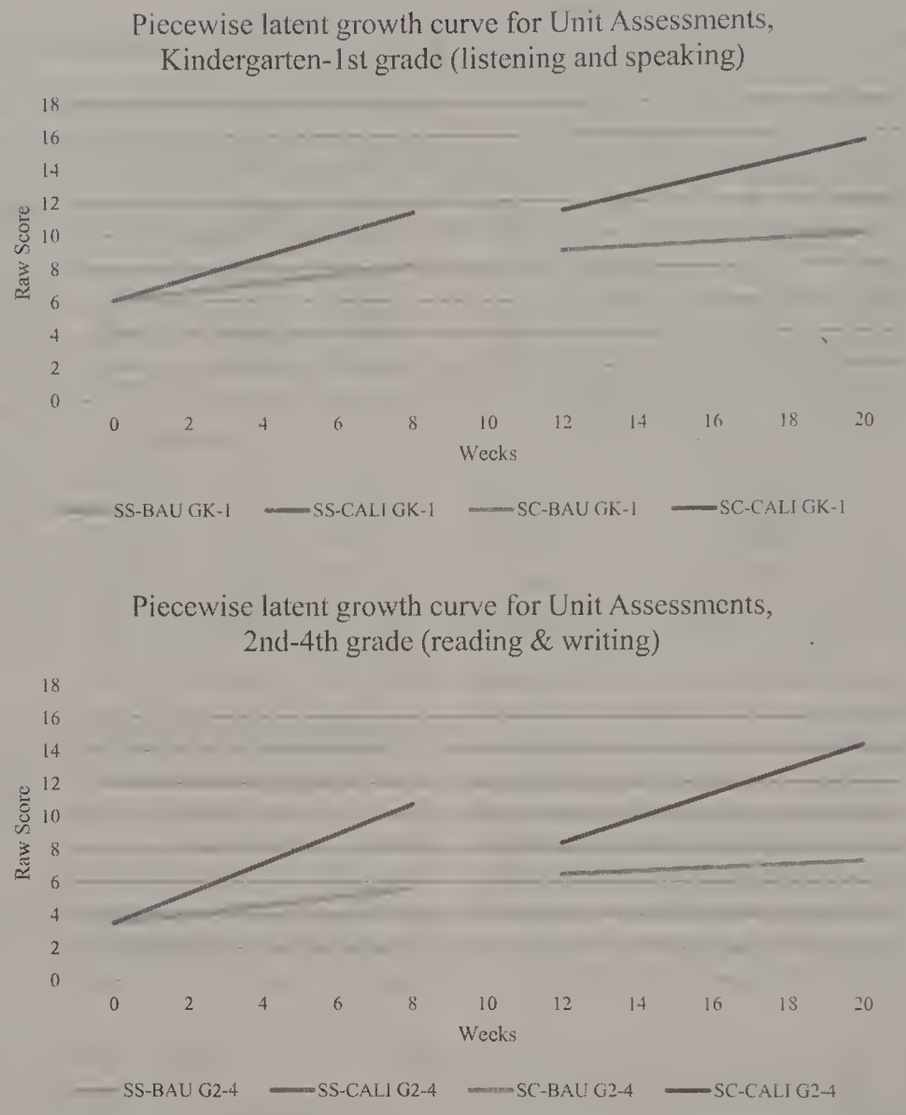


Figure 3. Model fitted results of piecewise hierarchical linear modeling growth curve models for kindergarten and first grade (top) and second through fourth grade (bottom). CALI SS was implemented during Weeks 0–8 and CALI SC was implemented during Weeks 12–20. SS = social studies; SC = science; BAU = business as usual; CALI = content-area literacy instruction; G = grade; K = kindergarten.

are provided in the online Supplemental Materials. We added all three preintervention standardized measures and interaction terms to the model at Level 2 and then, to preserve parsimony, trimmed nonsignificant effects. Final HLM results using our piecewise model revealed that oral comprehension was associated with performance on the unit assessment at the beginning of the study but not with growth (see Table S.2 and Figure 4 for model results). We found a C × I interaction effect for preintervention passage comprehension such that children who had higher initial passage comprehension scores made greater gains in CALI social studies than did children who had lower scores. The interaction effect reversed for CALI science, such that students with weaker preintervention passage comprehension scores made greater gains in science than did students with stronger scores. Hence, by the end of the four CALI units, the C × I interaction effects cancelled each other out.

Treatment effects on distal measures and assessing opportunity cost. To assess potential opportunity cost, we examined treatment effects—either positive or negative—for standardized assessment outcomes. We found no evidence of opportunity cost. Using HLM models with students nested in classrooms, we found a significant positive effect of treatment for Picture Vocabulary (d = 1.20), for Oral Comprehension (d = .47), and for Passage Comprehension (d = .22) for fourth graders. There was no effect of treatment—either positive or negative—in any other grades. For R2C, which was administered only to third and fourth graders, HLM multivariate multilevel models, with items nested in students nested in classrooms, revealed a positive effect of treatment, which was significantly greater for fourth graders than for third graders (see Table 7 and Figure 5).

Testing our theory of change. To test our theory of change (see Figure 1, bottom), we used structural equation modeling (AMOS version 22.0). The path diagram, with standardized path coefficients, is provided in Figure 6. In this model, CALI assignment predicts content knowledge using the last unit posttest, which takes into account the accumulation of content knowledge from the beginning to the end of the 12 weeks and four units. We then investigated the associations among our two language measures,

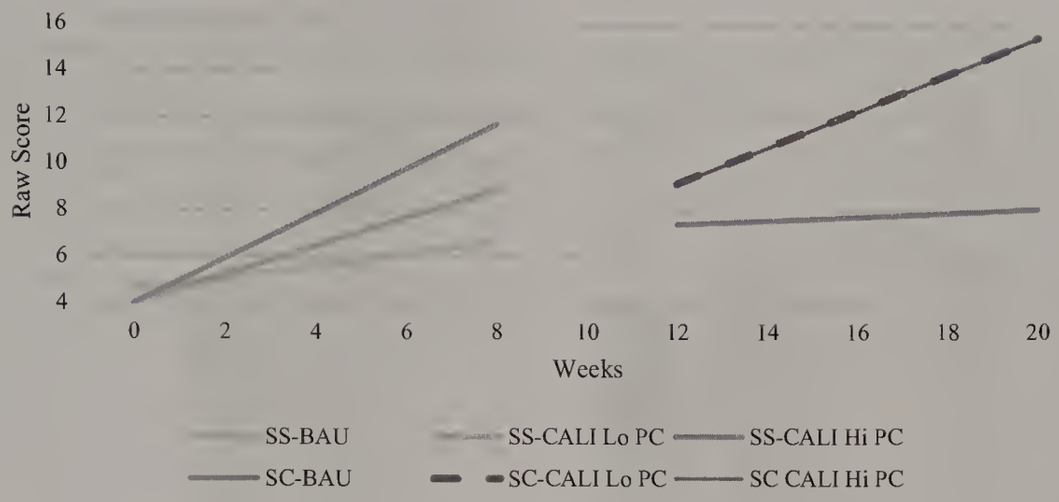


Figure 4. Model results comparing students who have preintervention passage comprehension scores one standard deviation (SD = 34) above the mean (high PC, M = 500) and below the mean (low PC, M = 432) for the sample. Treatment effect sizes (d) for social studies are 1.80 and .78, respectively. Treatment effect size (d) for science is 2.59 with no Child × Instruction interactions. BAU = business as usual; CALI = content-area literacy instruction; SS = social studies; SC = science; PC = passage comprehension. See the online article for the color version of this figure.

oral comprehension and vocabulary, and with the reading measure, passage comprehension. We only used assessments that were administered to all of the students from kindergarten through fourth grade. We used estimated marginal means to account for missing data. The fit of the data was adequate using three widely used fit indices (Hoyle, 1995): Tucker–Lewis index = .940, comparative fit index = .988, and root mean square error of approximation = .080, (p-close = .113).

Overall, our theory of change was supported (see Figure 6 and Table 8). Participation in CALI significantly predicted stronger performance on the last unit posttest score, which, in turn, predicted stronger vocabulary, oral comprehension and passage comprehension. Oral comprehension also predicted passage comprehension. The total standardized effect (direct plus indirect effects) of CALI participation on passage comprehension was 0.125. Overall, the model explained 50% of the variance in passage comprehension. The model assumes that oral comprehension predicts passage comprehension but the directions could be reversed because the standardized measures were assessed concurrently albeit after the unit posttest. Thus, there are plausible alternative models with similar fit and estimate results. For example, a model with spring vocabulary directly predicting comprehension, rather than oral comprehension, had identical fit and highly similar path coefficients. Reciprocal effects are also plausible but testing them was beyond the scope of this study.

Discussion

The results of our two studies reveal that CALI, when implemented with kindergarteners through fourth graders, can effectively improve students' social studies and science content knowledge, with large effects on proximal measures of social studies and science knowledge. During CALI lessons, students are provided with systematic opportunities to talk about, read, and experience social studies and science texts and content. This includes using original sources in social studies and experiments in science. Results also show that CALI can be an integral part of the literacy block without jeopardizing language and literacy learning with modest but positive effects on oral and reading comprehension skills (i.e., no opportunity cost). By the end of the school year, there were effects of treatment on reading comprehension and language skills. Specifically, third and fourth graders in the CALI condition demonstrated higher scores on the researcher developed R2C measure compared to students in the control group, and there was a small total effect on passage comprehension.

Table 7
Multilinear Multivariate Model of Reading-2-Comprehension for Third and Fourth Graders Where Outcomes are Expressed as Log-Odds

Fixed effect	Coefficient	*SE*	*t* ratio	Approximate *df*	*p* value
Log-odds for third graders	.350	.029	11.868	13	<.001
Effect for fourth graders	.021	.063	.336	13	.742
CALI effect for third graders	.041	.031	1.338	152	.181
CALI effect for fourth graders	.170	.065278	2.608	13	.022

Note. CALI = content-area literacy instruction. Deviance = 1,838.37. Third grade was the fixed reference group (= 0; fourth grade = 1).

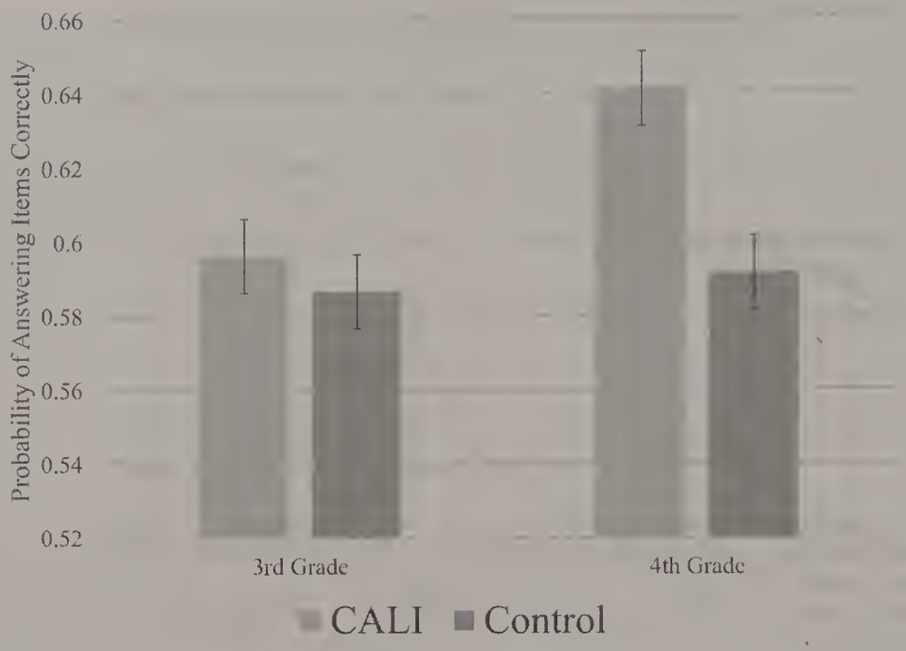


Figure 5. Probability of answering reading-to-comprehension items correctly as a function of condition and grade. CALI = content-area literacy instruction.

As can be seen in Figure 3, content knowledge gains accumulated over time as students completed each 3-week unit. That is, as students participated in CALI, their performance on the unit assessments increased, compared to the control group, until, at the final posttest, the effect of CALI was large by any standard. Moreover, there was some evidence of transfer when CALI shifted from social studies to science (see Figures 3 and 5), which was unexpected. We conjecture that instruction on how to respond to open-ended questions as well as strategies for reading expository text may have contributed to this transfer. It was the case, for example, that the number of words written in response to the open-ended questions increased from social studies Unit A (first unit, $M = 26$ words on three questions) to science Unit B (last unit, $M = 35$ words), $t(229) = 6.31$, $p < .001$, and, within content areas, from Unit A to Unit B (social studies, $M = 26$ to 28 words, $p = .047$; science, $M = 28$ to 35 words, $p < .001$). CALI units were designed to build on one another and so the accumulation of knowledge and transfer of literacy skills from social studies to science learning is highly encouraging.

Finally, we partially demonstrated that general education instruction in social studies and science could be effectively individualized (or personalized) for students, regardless of incoming background knowledge, language, and reading skills. There were less than ideal C × I interactions for social studies whereby students with weaker reading comprehension skills made smaller gains in social studies knowledge compared to students with stronger skills. However, this C × I interaction reversed for science—students with weaker initial reading comprehension

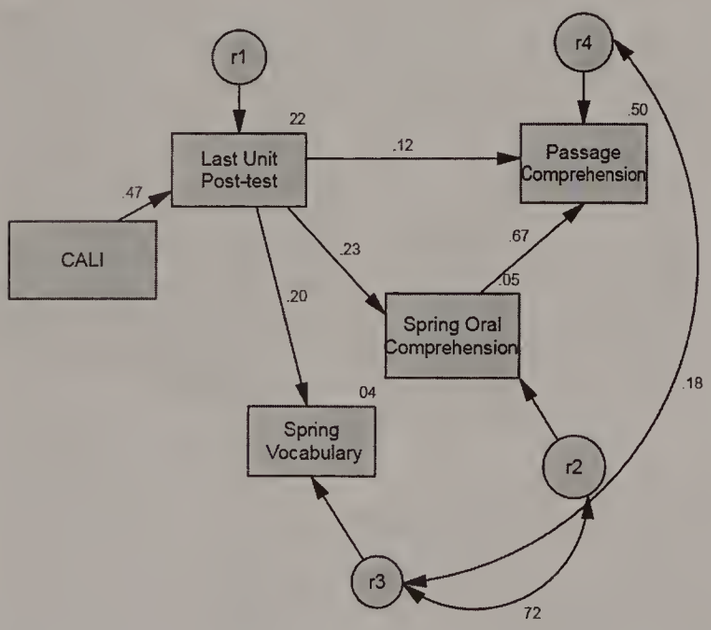


Figure 6. Structural equation model path diagram. Path coefficients are standardized. Numbers by variable boxes are variance explained. Curved lines are correlations. Fit was adequate: Tucker–Lewis index = .940, comparative fit index = .988, and root mean square error of approximation = .080 (p-close = .113). See the online article for the color version of this figure.

Table 8
Standardized Total Effects for Condition (Content-Area Literacy Instruction vs. Comparison Group) on Postintervention Unit Posttest Raw Score (RS), Oral Comprehension, Picture Vocabulary, and Passage Comprehension Developmental Scores (W)

Measure	Condition	Final unit posttest	Oral Comprehension W
Final unit posttest RS	.468		
Oral Comprehension W	.106	.226	
Picture Vocabulary W	.092	.195	
Passage Comprehension W	.125	.268	.670

skills made greater gains in science knowledge than did students with stronger skills. The net effect was similar outcomes regardless of incoming reading comprehension skill.

CALI was developed over multiple iterations using DBIR. In conducting DBIR, the procedures we used involved iterating between design and evaluation to continually refine the intervention toward the aims established by the theory of change (see Figure 1). The design team, which included researchers, teachers, and principals was critical to this process. Our design instances and small-t theories about how those designs functioned evolved over time, and were informed both by usability and outcome data generated through our iterative design pre–post studies with CALI. For instance, we learned that to reduce C × I interactions required multiple methods in combination, including leveled text, small flexible learning groups, and assessment-guided instruction.

Limitations

There are limitations that should be considered while interpreting these results. First, all of the schools in our studies were higher poverty schools; at the most affluent school, 41% of students received FARL (*M* across all schools = 57%). These results may not generalize to students attending more affluent schools. Second, the efficacy study was powered to find educationally important effects studywide but was arguably underpowered to find grade specific effects.

Third, when brought to scale, it is most likely that students' assignment to the CALI higher performing blue, average performing yellow, and lower performing green groups will change over the school year as students' skills change, as it did in the design studies. We lacked the resources to power this more dynamic protocol in the RCT, and hence the students spent the entire year in the same level group. The impact of CALI might differ if students change group level as their reading skills improve. Plus, the static groups may have contributed to C × I interactions because the RCT used static grouping whereas the design studies used dynamic grouping. The whole point of using flexible learning groups is that students' groups change as their skills and learning needs change. Finally, children in the average and lower performing groups were more likely to leave the district compared to students in the higher performing groups although the rate of attrition for students with weaker fall reading skills was the same for both treatment and control groups. This limitation should also be considered when interpreting the RCT results.

Child × Instruction Interactions

An explicit aim of CALI was to eliminate C × I interactions. A key reason that the DBIR required multiple iterations was because eliminating C × I interactions was difficult. Just as we solved one problem, another popped up. As one reviewer noted,

> It may be frustrating that solving one problem pops up another, but it also highlights how processes such as C × I interactions are contextualized . . . and should not be assumed to give way for simple solutions. This is an important theoretical and practice insight from [design studies] not a problem.

As noted above, we found C × I interactions in the RCT. Quantile regression results presented in the online Supplemental Materials (Figure S.2) showed some variation in treatment effects by quantile on the unit posttests, particularly at the tails. Taken together, these findings show that individual student's differential responses to instruction are pervasive and complicated to design against. We were generally successful in ensuring that, regardless of incoming skills, all students gained important content knowledge—but arguably, this was more the case with science and less with social studies. We attribute any success to DBIR and the important contributions of our teachers and partner principals. In addition, our findings offer insight into the challenges of designing effective instructional regimes and the importance of going beyond development and promise to the testing of efficacy. Of course, as we noted in the introduction, researchers conducting DBIR and RCTs have different aims and epistemological assumptions. We acknowledge the tensions and point out that the development and evaluation of CALI provides an important example of how competing frameworks can work synergistically to elucidate theory and practice (Creswell & Clark, 2011).

Given the very real challenges of developing CALI, we wonder why anybody would think that researchers should develop instructional programs without teachers and educational leaders on the design team, or that classroom teachers should develop their own curriculum lessons and materials without researcher support. Yet, many instructional programs and interventions are designed to answer research questions and to test theory rather than to actually create instructional regimes that are actionable in today's classrooms. Plus, it is a pervasive expectation among school and district administrators, as well as many teacher preparation programs, that teachers develop their own curriculum, lessons, and materials. To develop CALI as an efficacious instructional regime took 2 years of careful teamwork among teachers, researchers, and educational leaders, using research funding through the Reading for Understanding Network (RFU; U.S. Department of Education, Institute of Education Sciences). The RFU funding accelerated what would have, using other funding mechanisms, taken at least 7 years (3 for development and 4 more for efficacy). DBIR allowed us to fail fast and learn from mistakes; continuous RFU funding allowed us to move directly from development to efficacy—and to develop CALI science during the first half of the efficacy trial.

The next challenge will be to bring effective instructional programs to the classroom by increasing use of effective standards of practice for educational professionals, and reducing the current idiosyncratic practices that are pervasive (Raudenbush, 2005). School-research partnerships are an important and evolving practice (Coburn, Penuel, & Geil, 2013). Improving how we educate preservice and in-service teachers, educational leaders, and researchers, along with changing practitioners' and policymakers' perceptions about the importance of research and efficacy testing will be crucial next steps.

Opportunity Cost

A key assumption for why there is less time spent teaching social studies and science in the early grades is that time teaching them is time taken away from teaching reading (Williams et al., 2009). Our results suggest that this is not the case. Even with research teachers rather than classroom teachers implementing CALI, there was no opportunity cost. Students in CALI generally performed as well as or better on a standardized measure of

reading comprehension, compared to students in the control group. This was the case even considering that the classroom teacher had fewer students to teach during the literacy block while CALI was implemented. These results replicate and extend findings by Williams and colleagues (2009) for second graders to include kindergarteners through fourth graders. Teaching content-area literacy effectively does not preclude students' reading gains and, particularly for later grades, may enhance oral and written comprehension. Indeed, as we discuss next, our theory of change was supported, indicating that gaining content knowledge appears to be an important contributor to proficient reading for understanding.

Testing our Theory of Change

Although we started with a model based on the DIME model (see Figure 1, top), a simpler model emerged during the DBIR and was tested with the efficacy study data (see Figure 1, bottom, and Figure 6). In this model, we conjectured that improving content knowledge would impact both oral and written comprehension skills and that there would be direct effects of CALI on reading comprehension skills, as well as indirect effects through improved oral comprehension and vocabulary. Testing reciprocal effects was beyond the scope of this study. The elements of our theory of change that we could test were supported. Results revealed that increasing content knowledge was associated with gains in vocabulary, oral comprehension, and passage comprehension. In reviewing total effects (direct and indirect) of content knowledge on the constructs of interest (see Table 8), we found stronger total effects for oral and reading comprehension and smaller effects for vocabulary.

The emerging constructs of academic knowledge and academic language (Snow, 2010) are supported by the model—content knowledge predicts both vocabulary and oral comprehension, as well as reading comprehension (see Figure 6). Academic language is defined as the more formal language that is used increasingly in classrooms as students' schooling progresses into middle school and beyond. It requires specialized vocabulary and a sophisticated semantic system, a more formal syntax, strong world and content knowledge, and metacognition. The Common Core Standards and the Next Generation Science Standards, for example, mandate critical thinking and making inferences across texts, which requires strong academic language as well as metacognition. The increasing development of metacognition, which emerges around the ages of 8 and 9 years, may help to explain the greater effect sizes in fourth grade (see Table 6).

In sum, our findings reveal that kindergarten through fourth grade is not too early to teach content-area literacy, which can be taught during the dedicated block of time devoted to reading instruction without negatively impacting reading gains. We did observe C × I interactions in social studies (students with weaker reading comprehension skills made weaker gains compared to those with stronger skills) as well as science although the C × I interactions had the opposite effect for science (students with weaker comprehension made greater gains compared to those with stronger skills). Hence, continued focus on individual differences among students and the importance of dynamically individualizing (or personalizing) the instruction they receive, based on their developing skills, will help meet the aims of all students achieving their potential and mitigating achievement gaps. Content-area literacy instruction, which takes into account individual student differences, is explicit and systematic, and encourages students to read, write about, and talk about expository text, while also incorporating disciplinary practices, can effectively increase students' development of content knowledge through the early, middle, and later elementary grades, with positive effects on oral and written comprehension.

References

Alleman, J., & Brophy, J. (2003). History is alive: Teaching young children about changes over time. *Social Studies, 94,* 107–110. http://dx.doi.org/10.1080/00377990309600191

Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., Meadows, J., & Li, Z. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *Elementary School Journal, 111,* 535–560.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Educational Research Journal, 14,* 367–381. http://dx.doi.org/10.3102/00028312014004367

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher, 41,* 16–25. http://dx.doi.org/10.3102/0013189X11428813

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., & Campbell, A. M. (2013). *Report of the 2012 national survey of science and mathematics education.* Chapel Hill, NC: Horizon Research, Inc.

Blank, R. K. (2012). *What is the impact of decline in science instructional time in elementary school? Time for elementary instruction has declined, and less time for science is correlated with lower scores on NAEP* Paper prepared for the Noyce Foundation. Retrieved from www.csssscience.org/downloads/NAEPElemScienceData.pdf

Blumenfeld, P., Fishman, B., Krajcik, J. S., Marx, R. W., & Soloway, E. (2000). Creating usable innovations in systemic reform: Scaling-up technology-embedded project-based science in urban schools. *Educational Psychologist, 35,* 149–164. http://dx.doi.org/10.1207/S15326985EP3503_2

Brophy, J., Alleman, J., & O'Mahony, C. (2003). Primary-grade students' knowledge and thinking about food production and the origins of common foods. *Theory and Research in Social Education, 31,* 10–49. http://dx.doi.org/10.1080/00933104.2003.10473214

Bybee, R. W. (1997). *Achieving science literacy: From purposes to practice.* Portsmouth, NH: Heinemann.

CCS Common Core State Standards Initiative. (2010). Common Core state standards for mathematics. Retrieved March 2013, from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Chi, M. T. H., Fletovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 6,* 121–152. http://dx.doi.org/10.1207/s15516709cog0502_2

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 257–273. http://dx.doi.org/10.1016/S0022-5371(79)90146-4

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32,* 9–13. http://dx.doi.org/10.3102/0013189X032001009

Coburn, C. E., Penuel, W. R., & Geil, K. E. (2013). Research-practice partnerships: A strategy for leveraging research for educational improvement in school districts. Retrieved from http://forumfyi.org/files/R-P%20Partnerships%20White%20Paper%20%20Jan%202013%20-%20Coburn%20Penuel%20&%20Geil.pdf

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25,* 119–142. http://dx.doi.org/10.3102/01623737025002119

Connor, C. M., Kaya, S., Luck, M., Toste, J., Canto, A., Rice, D. C., . . . Underwood, P. (2010). Content-area literacy: Individualizing student instruction in second grade science. *The Reading Teacher, 63,* 474–485. http://dx.doi.org/10.1598/RT.63.6.4

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science, 24,* 1408–1419. http://dx.doi.org/10.1177/0956797612472204

Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J., Lundblom, E., Crowe, E. C., & Fishman, B. (2011). Effective classroom instruction: Implications of child characteristic by instruction interactions on first graders' word reading achievement. *Journal of Research on Educational Effectiveness, 4,* 173–207. http://dx.doi.org/10.1080/19345747.2010.510179

Connor, C. M., Phillips, B. M., Kaschak, M., Apel, K., Kim, Y.-S., Al Otaiba, S., . . . Lonigan, C. J. (2014). Comprehension tools for teachers: Reading for understanding from prekindergarten through fourth grade. *Educational Psychology Review, 26,* 379–401. http://dx.doi.org/10.1007/s10648-014-9267-1

Connor, C. M., Rice, D. C., Canto, A. I., Southerland, S. A., Underwood, P., Kaya, S., . . . Morrison, F. J. (2012). Child characteristics by science instruction interactions in second and third grade and their relation to students' content-area knowledge, vocabulary, and reading skill gains. *The Elementary School Journal, 113,* 52–75. http://dx.doi.org/10.1086/665815

Creswell, J. W., & Clark, V. P. (2011). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation (DIME) model of reading comprehension. *Journal of Educational Psychology, 99,* 311–325. http://dx.doi.org/10.1037/0022-0663.99.2.311

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching, 37,* 582–601. http://dx.doi.org/10.1002/1098-2736(200008)37:6<582::AID-TEA5>3.0.CO;2-L

Diakidoy, I. N., & Kendeou, P. (2001). Facilitating conceptual change in astronomy: A comparison of the effectiveness of two instructional approaches. *Learning and Instruction, 11,* 1–20. http://dx.doi.org/10.1016/S0959-4752(00)00011-6

Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69,* 145–186. http://dx.doi.org/10.3102/00346543069002145

Donovan, M. S. (2013). Generating improvement through research and development in education systems. *Science, 340,* 317–319. http://dx.doi.org/10.1126/science.1236180

Duke, N. K. (2000). 3.6 minutes per day: The scarcity of informational texts in first grade. *Reading Research Quarterly, 35,* 202–224. http://dx.doi.org/10.1598/RRQ.35.2.1

Fishman, B., Marx, R., Blumenfeld, P., Krajcik, J. S., & Soloway, E. (2004). Creating a framework for research on systemic technology innovations. *Journal of the Learning Sciences, 13,* 43–76. http://dx.doi.org/10.1207/s15327809jls1301_3

Fishman, B. J., Penuel, W. R., Allen, A.-R., Cheng, B. H., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *Yearbook of the National Society for the Study of Education, 112,* 136–156.

Fitchett, P. G., Heafner, T. L., & Lambert, R. G. (2010). *Social studies under siege: Examining policy and teacher-level factors associated with elementary social studies marginalization.* Charlotte, NC: University of North Carolina.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First Impact Study Final Report (NCEE 2009–4038).* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Graham, S., Harris, K. R., & Hebert, M. A. (2011). It is more than just the message: Presentation effects in scoring writing. *Focus on Exceptional Children, 11,* 1–12.

Grant, M. C., & Fisher, D. B. (2010). *Reading and writing in science: Tools to develop disciplinary literacy.* Thousand Oaks, CA: Corwin.

Grant, S. G., & Salinas, C. (2008). Assessment and accountability in the social studies. In L. S. Levstik & C. A. Tyson (Eds.), *Handbook of research on social studies education.* New York, NY: Routledge.

Hirsh, E. D. (2006). *The knowledge deficit: Closing the shocking education gap for American children.* New York, NY: Houghton Mifflin.

Hoge, J. D. (1996). *Effective elementary social studies.* Stamford, CT: Wadsworth Publishing.

Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications.* Thousand Oaks, CA: Sage.

Jeong, J., Gaffney, J. S., & Choi, J. O. (2010). Availability and use of informational texts in second-, third-, and fourth-grade classrooms. *Research in the Teaching of English, 44,* 435–456.

Kendeou, P., & van den Broek, P. (2005). The effects of readers' misconceptions on comprehension of scientific text. *Journal of Educational Psychology, 97,* 235–245. http://dx.doi.org/10.1037/0022-0663.97.2.235

Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica, 46,* 33–50. http://dx.doi.org/10.2307/1913643

Lawrence Hall of Science. (2007). *Seeds of science/roots of reading.* Nashua, NH: Delta Education LLC & Regents of the University of California.

Macken, C. (2003). What in the world do second graders know about geography? Using picture books to teach geography. *Social Studies, 94,* 63–68. http://dx.doi.org/10.1080/00377990309600184

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual: Woodcock-Johnson III.* Itasca, IL: Riverside Publishing.

McMurrer, J. (2008). *Instructional time in elementary schools: A closer look at changes for specific subjects.* Washington, DC: Center on Education Policy.

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Science achievement gaps begin early, persist, and are largely explained by modifiable factors. *Educational Researcher, 45,* 18–35.

National Assessment of Educational Progress. (2011). The nation's report card. Retrieved from http://nces.ed.gov/nationsreportcard/

National Council for the Social Studies (NCSS). (1994). *Curriculum standards for social studies: Expectations of excellence.* Washington, DC: National Council for the Social Studies. Retrieved from http://www.socialstudies.org/standards/introduction

National Science Education Standards. (1996). *Front matter.* Washington, DC: The National Academies Press.

Next Generation Science Standards. (2013). Appendix M: Connections to the Common Core state standards for literacy in science and technical subjects. Retrieved from http://www.nextgenscience.org/sites/default/files/Appendix%20M%20Connections%20to%20the%20CCSS%20for%20Literacy_061213.pdf

Palincsar, A. S., & Magnusson, S. J. (2001). The interplay of first-hand and text-based investigations to model and support the development of scientific knowledge and reasoning. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty five years of progress* (pp. 151–194). Mahwah, NJ: Erlbaum.

Pearson, P. D., Moje, E., & Greenleaf, C. (2010). Literacy and science: Each in the service of the other. *Science, 328,* 459–463. http://dx.doi.org/10.1126/science.1182595

Petscher, Y., & Logan, J. A. (2014). Quantile regression in the study of developmental sciences. *Child Development, 85,* 861–881. http://dx.doi.org/10.1111/cdev.12190

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34,* 25–31. http://dx.doi.org/10.3102/0013189X034005025

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rawson, K. A., & Kintsch, W. (2002). How does background information improve memory for text content? *Memory & Cognition, 30,* 768–778. http://dx.doi.org/10.3758/BF03196432

Rawson, K. A., & Kintsch, W. (2004). Exploring encoding and retrieval effects of background information on text memory. *Discourse Processes, 38,* 323–344. http://dx.doi.org/10.1207/s15326950dp3803_3

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology, 80,* 16–20. http://dx.doi.org/10.1037/0022-0663.80.1.16

Risinger, C. F., & Garcia, J. (1995). National assessment and the social studies. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 68,* 225–228. http://dx.doi.org/10.1080/00098655.1995.9957237

Romance, N. R., & Vitale, M. R. (2001). Implementing an in-depth expanded science model in elementary schools: Multi-year findings, research issues, and policy implications. *International Journal of Science Education, 23,* 373–404. http://dx.doi.org/10.1080/09500690116738

Shadish, W. R., Cook, T. D., & Campbell, J. R. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York, NY: Houghton Mifflin Company.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science, 328,* 450–452. http://dx.doi.org/10.1126/science.1182597

Snow, C. E., Lawrence, J. F., & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness, 2,* 325–344. http://dx.doi.org/10.1080/19345740903167042

Voss, J. F., Fincher-Kiefer, R. H., Greene, T. R., & Post, T. A. (1986). Individual differences in performance: The contrastive approach to knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence.* Hillsdale, NJ: Erlbaum.

Wharton-McDonald, R., Pressley, M., & Hampston, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and student achievement. *The Elementary School Journal, 99,* 101–128. http://dx.doi.org/10.1086/461918

Williams, J. P., Stafford, K. B., Lauer, K. D., Hall, K. M., & Pollini, S. (2009). Embedding reading comprehension training in content-area instruction. *Journal of Educational Psychology, 101,* 1–20. http://dx.doi.org/10.1037/a0013152

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson-III Tests of Achievement.* Itasca, IL: Riverside.

Received June 17, 2015
Revision received March 21, 2016
Accepted March 21, 2016 ■

Correction to Glaser and Schwan (2015)

In the article "Explaining Pictures: How Verbal Cues Influence Processing of Pictorial Learning Material" by Manuela Glaser and Stephan Schwan (*Journal of Educational Psychology*, 2015, Vol. 107, No. 4, 1006–1018. http://dx.doi.org/10.1037/edu0000044), there were several errors in the **Results** section. All of the η^2 values should have been η_p^2 values.

http://dx.doi.org/10.1037/edu0000173

Journal of Educational Psychology
2017, Vol. 109, No. 3, 321–337

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000143

The Effects of Explicit Teaching of Strategies, Second-Order Concepts, and Epistemological Underpinnings on Students' Ability to Reason Causally in History

Gerhard L. Stoel, Jannet P. van Drie, and Carla A. M. van Boxtel
University of Amsterdam

This article reports an experimental study on the effects of explicit teaching on 11th grade students' ability to reason causally in history. Underpinned by the model of domain learning, explicit teaching is conceptualized as multidimensional, focusing on strategies and second-order concepts to generate and verbalize causal explanations and epistemological underpinnings connected to causal reasoning in history. In a randomized pretest–posttest design (N = 95), with a treatment and a control condition, effects of explicit teaching were investigated on students' (a) second-order and strategy knowledge, (b) their epistemological beliefs, and (c) their ability to construct a causal explanation, as well as (d) their topic knowledge, and (e) their individual interest. Results show that students in the experimental group scored significantly higher at the posttest on knowledge of causal-reasoning strategies and second-order concepts (sr^2 = .09), attributed a significantly higher value to criterialist epistemological beliefs (sr^2 = .04), and reported a higher individual interest (sr^2 = .02). We found no differences between conditions in the overall quality of students' written explanations. However, the experimental group scored significantly higher on 1 core criterion, that is, the "use of second-order language and causal connections" (sr^2 = .06). No differences were found on first-order knowledge. Furthermore, self-reports on learning gains and correlational analysis were applied to explore the interrelatedness of second-order and strategy knowledge, epistemological beliefs, student's ability to construct a causal explanation, topic knowledge, and individual interest.

Keywords: historical reasoning, history instruction, explicit teaching, instructional design, epistemological beliefs

Over the past two decades, researchers of history education have emphasized the importance of history education as a subject that allows students to develop skills and competencies which are considered important in a democratic and pluralistic society (Barton & Levstik, 2004). As a consequence, historical reasoning has been included in recent years in the national history curricula in many countries (e.g., the Netherlands, Australia, Canada, the United Kingdom; Erdmann & Hassberg, 2011). Among other things, students should learn to reason critically with and about multiple sources; to judge the reliability, usefulness, and representativeness of these sources; and to embed them in their historical context. Furthermore, students should learn to construct and deconstruct historical narratives, which demands understanding that these narratives do not primarily present copies of the past but interpretations and that multiple perspectives can coexist. Finally, students should learn to judge the validity of these interpretations using disciplinary criteria (Seixas & Morton, 2013; VanSledright, 2011; Wineburg, 2001).

Although the importance of teaching historical reasoning skills is widely accepted, relatively little is known about pedagogical principles that foster the development of this reasoning (Levstik & Barton, 2008; van Boxtel & van Drie, 2013). In a previous literature review and an experimental pilot study, several basic principles of a learning environment, intended to foster causal historical reasoning, were defined (i.e., designing open-ended tasks, allowing for social interaction, raising situational interest; Stoel, van Drie, & van Boxtel, 2015). However, our review and pilot study also showed the indispensability of explicit teaching as a design principle in a learning environment intended to develop students' ability to reason causally in history.

Previous studies have shown the effectiveness of explicit teaching in distinct topics such as sourcing strategies in history (Nokes, Dole, & Hacker, 2007; Reisman, 2012), writing historical essays (De La Paz, 2005; De La Paz & Felton, 2010), and epistemological beliefs in science (Khishfe & Abd-El-Khalick, 2002). The current study adds to this research by conceptualizing the focus of explicit teaching in a more integral fashion. We argue from both a theoretical and an empirical standpoint that fostering a causal historical reasoning skill entails teaching explicitly about the strategies and

This article was published Online First June 20, 2016.

Gerhard L. Stoel, Jannet P. van Drie, and Carla A. M. van Boxtel, Research Institute of Child Development and Education, University of Amsterdam.

We kindly thank the history teachers and students from the Keizer Karel College, Amstelveen, the Netherlands, for their involvement. We also thank Terrence Jorgensen, University of Amsterdam, for his help with the statistical analyses.

Correspondence concerning this article should be addressed to Gerhard L. Stoel, Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS, Amsterdam, the Netherlands. E-mail: g.l.stoel@uva.nl

second-order concepts related to historical causation and about the epistemological underpinnings of constructing historical explanations. To investigate the effects of explicit teaching on students' causal-historical reasoning, a randomized controlled trial with two conditions was conducted with 95 eleventh grade students.

Developing Causal Historical Reasoning

Prior to defining pedagogical principles, we conducted a literature review to delineate the cognitive dimensions involved in developing causal-historical reasoning. The model of domain learning (MDL; Alexander, 2003, 2005) provided an appropriate framework toward this goal. Elaborating on this model, we differentiated between (a) knowledge of causal strategies and second-order concepts related to historical causation and (b) epistemological beliefs about the nature of causal interpretations in history, as important aspects underpinning causal historical reasoning. Besides, the MDL conceptualized first-order knowledge and (situational) interest to be important ingredients of this reasoning (Stoel et al., 2015).

The MDL emphasizes that developing expertise in any domain involves acquiring domain-specific, deep-level strategies. These strategies allow a student to construct or critically evaluate new information in ways that are accepted within the given discipline. Within the context of causal-historical reasoning, important strategies to master are, among others, (a) to look for multiple causes; (b) to construct complex—as opposed to simple linear—causal models; (c) to analyze causes along multiple dimensions such as time, content, role; and (d) to analyze individuals' motives and actions in the context of the broader political, economic, cultural, and social context of the time (Chapman, 2003; Coffin, 2004; Halldén, 1997; Seixas & Morton, 2013). In addition to these strategies, students need to develop their knowledge of the second-order concepts, which historians use to construct causal narratives about the past (e.g., categorizing causes requires concepts such as direct, indirect, long term, short term, trigger, catalyst, precondition; contextualizing motives and actions requires concepts such as cultural, political). These second-order concepts give students the vocabulary to verbalize their causal reasoning and, more important, provide them with the conceptual apparatus to reason causally in history (i.e., to engage in deep-level strategies; van Drie & van Boxtel, 2008; VanSledright & Limón, 2006; Woodcock, 2005).

Another aspect that should be addressed in a learning environment aiming at fostering expertise are students' beliefs about the "complexity, sophistication and uncertainty of knowledge" (Alexander, 2005, p. 38). Alexander (2005) stated that students with more nuanced epistemological beliefs "tend to be higher academic achievers, report more strategic processing, and are more persistent in the face of difficulty" (p. 38). VanSledright and Limón (2006) suggested that epistemological beliefs and historical understanding are linked and that teaching historical reasoning involves influencing epistemological beliefs.

Within the field history education, epistemological beliefs have often been conceptualized in three "stances," *copier*, *subjectivist*, and *criterialist*. This stage model is embedded in more general theories about epistemological beliefs (see King & Kitchener, 2002; Kuhn & Weinstock, 2002; Maggioni, Alexander, & VanSledright, 2004; Maggioni, VanSledright, & Alexander, 2009). Operationalized for causal reasoning, students with a copier stance would believe that historical explanations should be a "copy" of the past and that inconclusive or contradictory evidence makes writing history impossible. In this stance, little value is placed on methodology because explanations are either correct or wrong. Students with a subjectivist stance accept the fact that historical explanations are interpretations but lack an understanding and appreciation of the (academic) criteria to judge these interpretations. Often this stance leads to the belief that history is merely a matter of opinion. Only with a criterialist stance do students understand and appreciate both the constructed nature of historical explanations as well as the academic criteria for evaluating these causal statements (cf. Lee & Shemilt, 2009). Based on the theories of Alexander (2005) and VanSledright and Limón (2006), a positive relationship is expected between students' epistemological beliefs, their conceptual and strategy knowledge, and the quality of their historical reasoning.

The MDL emphasizes that students in the early phases of expertise development often rely on the use of generic, surface-level strategies (e.g., rereading, summarizing) when confronted with a problem in a specific domain, whereas experts tend to engage in domain specific, deep-level strategies (e.g., using second-order concepts to categorize causes and embedding the analysis in a broader political, economic, cultural, or social context). Furthermore, the MDL links nuanced epistemological beliefs to higher levels of strategic processing. Therefore, developing expertise in causal-historical reasoning is defined in this study as the acquisition of deep-level strategies and second-order concepts while simultaneously stimulating development of more nuanced ideas on the nature of historical knowledge and the criteria for evaluating and constructing historical explanations.

The Role of Individual and Situational Interest

The MDL conceptualizes interest as an important precondition for developing and engaging students in effortful domain-specific, deep-level strategies. The model differentiates between two types of interest: *situational interest* and *individual interest*. Individual interest can be defined as a relatively stable learner characteristic expected to gradually increase as a student gains more knowledge of the domain and the specific strategies and questions involved. As expertise develops, it becomes easier for a learner to connect new information to the broader domain and to prior knowledge and interests, thus, intrinsic motivation increases. In contrast, learners in the early phases of expertise rely on the teacher and the characteristics of the learning environment in order to increase their situational interest and to connect a new topic to the broader domain as well as to their prior knowledge and interests. Situational interest consists of valuing the relevance of what is to be learned and enjoying the learning activities (Alexander, 2003).

The Role of Explicit Teaching

In a previous literature review and in an experimental pilot study, several basic principles of a learning environment intended to foster causal-historical reasoning were defined (i.e., designing open-ended tasks, allowing for social interaction, and raising situational interest; Alexander, 2005; Collins, Brown, & Holum, 1991; Stoel et al., 2015). However, the MDL also maintains that deep-level strategies "cannot be expected to develop naturally but must be cultivated" (Alexander, 2005, p. 40).

This statement is corroborated by several studies in history education that focused on explicit teaching strategies related to analyzing sources—sometimes in combination with writing strategies (De La Paz, 2005; De La Paz & Felton, 2010; Nokes et al., 2007; Reisman, 2012). In these studies, positive effects were found on the quality and length of students' essays, the use of strategies, and general historical thinking. However, these studies all focused on reasoning with historical sources. No previous studies have focused on causal-historical reasoning. Furthermore, these studies had a quasi-experimental design and limited explicit teaching to instructing strategies.

This study adds to the current research by its randomized-controlled design and by focusing on causal-historical reasoning. The study also expands explicit teaching to include causal-historical strategies and second-order concepts as well as epistemological beliefs. In line with the MDL, effects are analyzed not only by measuring a complex causal historical skill, in the form of an essay task, but also by assessing the underlying aspects of causal-historical reasoning: knowledge of causal strategies, second-order concepts, and epistemological ideas. Including students' epistemological beliefs as a dependent variable is also advocated by Reisman (2012). Because of the centrality of first-order knowledge and individual interest in the MDL, these aspects are also measured.

To summarize, we designed this study to investigate the differences between a condition in which students work together on an open-ended explanatory task, while being explicitly taught about the concepts and strategies involved in causal reasoning and while reflecting on the epistemological aspects of their explanations (*explicit condition*), and a condition in which students work together on the same task but without this explicit teaching of strategies, second-order concepts, and epistemological underpinnings (*implicit condition*).

Research Question

Our central research question is as follows: What is the effect of explicit teaching on second-order concepts, causal reasoning strategies, and epistemological underpinnings (in the context of a collaborative explanatory task) on 11th grade students' (a) second-order and strategy knowledge, (b) their epistemological beliefs and (c) their ability to construct a causal explanation, compared with a control group working on a similar task without explicit attention to causal strategies, concepts, and epistemological beliefs? In addition, the effects of the teaching condition (explicit vs. implicit) on students' first-order knowledge and individual interest was compared.

The study was designed as a pretest–posttest randomized controlled experiment. In our analysis, we not only investigated the effects of explicit teaching on students' knowledge, beliefs, and skills, but we also explored the relationships between these different constructs at the pretest and posttest.

Hypotheses

Based on the theoretical framework, we hypothesized that the explicit teaching of causal reasoning strategies and second-order concepts and epistemological reflection—embedded in an open-ended, explanatory task (explicit condition)—would be an effective learning environment for fostering causal-historical reasoning, compared with a control condition that worked on the same task but without the explicit teaching (implicit condition). While controlling for differences in pretest scores and situational interest, we formed the following hypotheses regarding what would occur at posttest:

Hypothesis 1: In the explicit condition, knowledge of causal reasoning strategies and second-order concepts will be significantly higher, compared with the implicit condition.

Hypothesis 2: A significantly more nuanced epistemological stance (indicated by lesser agreement with subjectivist items and more in accord with criterialist items) will have developed in the explicit condition, compared with the implicit condition.

Hypothesis 3: The ability to construct a causal-historical explanation will be significantly higher in the explicit condition, compared with the implicit condition.

Hypothesis 4: Historical first-order knowledge will not differ between conditions

Hypothesis 5: Individual interest will not differ between conditions.

Method

Participants

In total, 104 eleventh grade preuniversity students from four history classes and two teachers participated in the experiment. The average age of the students was 16.8 years (minimum 16, maximum 19). In the Netherlands, preuniversity education (VWO) is the highest educational track in secondary education. Approximately 20% of the secondary school students are enrolled in this 6-year program (Grades 7 to 12). A preuniversity diploma allows admission to university. The participating school is a public school for higher general secondary and preuniversity education. The school has 1,700 students and is situated in a relatively prosperous, suburban community near Amsterdam—average income is 15% above national average (Central Bureau for Statistics, 2015). The lesson table is comprised of single or double 45-min units. History is a mandatory subject in two of the four predefined profiles from which students choose after the 9th grade. At this school, students receive three history lessons a week. World War I has previously been studied in Grade 9. Our lesson-unit marked the beginning of a module on the "Time of Two World Wars," one of the era's in the framework of orientation knowledge.

Within each of the four classes, students were randomly assigned to a condition, creating four experimental and four control subgroups (see Table 1). The subgroups could not be mixed across classes due to different timetables. Because we wanted students to work in triads, we made minor adjustments to the sample size in the subgroups to ensure that the number of students in the treatment subgroups was divisible by three. (When necessary, we added one or two students per class to the experimental subgroup—which explains the different sample size of the two conditions.) The subgroups from each class were inspected on (a) gender distribution and (b) average achievement, based on stu-

Table 1
Sampling Design

Class (class teacher) and condition	External teacher	*n* (students)
Class 1 (Teacher A)		
Explicit	1	15
Implicit	2	12
Class 2 (Teacher A)		
Explicit	2	12
Implicit	1	11
Class 3 (Teacher B)		
Explicit	1	15
Implicit	2	13
Class 4 (Teacher B)		
Explicit	2	15
Implicit	1	11

dents' history grades during the school year. This led to some minor exchanges between the subgroups per class. Subsequently, 19 triads were created in the four experimental subgroups. In the control condition, students worked in triads as well, but we allowed for one or two dyads to exist in each of the four subgroups. In total 13 triads and 4 dyads were created in the control condition. All triads were composed of a high scoring, a low scoring, and an average student (based on students' history grades during the school year) to prevent the confounding of outcomes on dependent variables with differences between triads.

One student opted not to participate before the experiment (n = 1). After the experiment, we excluded eight students who missed more than one intervention lesson (n_{exp} = 3; n_{imp} = 5). This resulted in a final sample size of 95 students. In our analysis, the explicit condition consisted of 53 students (28 male and 25 female) and the implicit condition consisted of 42 students (21 male and 21 female).

During the lesson-unit, the subgroups from each class worked in two separate classrooms and two external teachers instructed the groups. The first external teacher holds a degree in history and teaching and has taught history at the secondary level for 8 years. The second external teacher holds a PhD in history and a degree in teaching and has taught history at the secondary level for 2 years. We choose two external teachers to teach the intervention lessons—instead of the regular class teachers—in order to prevent the confounding of outcomes with potential teacher effects. To prevent differences between the external teachers (possibly leading to confound learning outcomes), the external teachers switched conditions between classes so that each taught two experimental and two control subgroups. Table 1 presents an overview of the sampling design.

Lesson-Unit and Procedures

The lesson-unit focused on explaining the outbreak of World War I. Before the start of the experiment, the students were tested on (a) knowledge of second-order concepts and causal-reasoning strategies, (b) epistemological beliefs (subjectivist and criterialist), (c) first-order knowledge, and (d) individual interest in history (see Table 2; Pretest I). Subsequently, all students received two preparatory lessons, both lasting 45 min, that focused on developing students' historical knowledge about events, countries, developments, and phenomena in Europe in the run-up to World War I (see Table 2; Lesson 1/2). With this preparation, we sought to provide students with enough first-order knowledge to reduce the confounding of students' reasoning abilities in the pretest essay-writing task with a low level of knowledge about the topic.

After the preparatory lessons, students wrote a history essay (pretest) using several sources and their prior knowledge to explain why Germany became involved in World War I (see Table 2; Pretest II). Subsequently, the actual intervention took place (see Table 2; Lessons 3, 4, & 5). The subgroups (explicit and implicit) worked for three consecutive lessons in separate classrooms on an open-ended collaborative task. At the end of the third lesson, students filled out a short questionnaire, designed to measure their situational interest in the previous lessons. After the experiment, students took one lesson to rewrite their pretest essays (see Table 2; Posttest I). Finally, students retook the tests that measured knowledge, epistemological beliefs, and individual interest (see Table 2; Posttest II).

The experiment was conducted during 2 weeks in March 2014. Pretest I was taken 4 weeks before the start of the intervention. The preparatory lessons and Pretest II took place in the first week. The intervention lesson, as well as Posttest I and II, took place in the second week. Table 2 presents a schematic summary of the design and the measurement instruments.

Designing the Implicit and Explicit Condition

Hereunder both lesson-units are described. Because the basic model and design principals were the same for both groups, we start with the commonalities and then move on to the elaboration

Table 2
Procedure

Lesson	Phase	Measurement
Before preparatory lessons	Pretest I	Knowledge: conceptual & strategic, first order Beliefs: epistemology Interest: individual
Lesson 1 and 2	Preparatory topic lessons	
Before intervention lessons	Pretest II	Skills: essay task
Lesson 3, 4, and 5	Intervention lessons (subgroups work in separate classrooms)	Interest: situational
After intervention lessons	Posttest I	Skills: rewriting essay task
	Posttest II	Knowledge: conceptual & strategic, first order Beliefs: epistemology Interest: individual

of the lesson-units in the two conditions. Table 3 provides a summary of both lesson-units. A detailed elaboration of the lesson goals can be found in Appendix A.

Commonalities. Both lesson-units were designed from a constructivist perspective on learning and followed characteristics of a problem based learning environment (Savery & Duffy, 1995) and pedagogical principles described in the MDL (Alexander, 2005) and the model of cognitive apprenticeship (Collins et al., 1991). Based on Merrill's (2002) review of the common characteristics of different instructional theories, we discerned four phases in the learning environment: a preparatory phase, an instructional phase, a phase of application, and a phase of integration.

In both lesson-units, students worked on an authentic, open-ended task, based on an exemplary question in the domain of history: "How can we explain the outbreak of the First World War?" Group work (triads) and whole class discussion was used to stimulate interaction and argumentation. An important characteristic of all learning activities was the aim to make students' thinking visible, thereby allowing the teacher to provide students with scaffolding, coaching, and constructive feedback.

In the preparatory phase, effort was made to raise situational interest and to allow students to understand the relevance of the topic by triggering prior knowledge and interest and connecting the topic to the broader domain. In the instructional phase, the key question and the task were explained and the goals of the lessons were explicated. During the phase of application, students in both conditions worked in triads to coconstruct an explanation. Students worked on card-sorting tasks and graphical representation to select, organize, and connect causes and to construct their explanations. Research has shown that graphical representations allow students to externalize and explicate their thinking, mediate their analysis and discussion, and enhance their historical thinking (Prangsma, van Boxtel, & Kanselaar, 2008; van Drie & van Boxtel, 2003; van Drie, van Boxtel, Jaspers, & Kanselaar, 2005). In the phase of integration, students presented their conclusions and discussed them in a whole-class setting. This phase intended to broaden and deepen students' thinking by allowing them to compare their explanations and to reflect on similarities and differences between their products. Furthermore, these whole-class discussions allowed the teacher to uncover and address possible misconceptions. Previous studies in history education have underscored the effectiveness of whole-class discussions (Havekes, 2015; van Drie & van Boxtel, 2011). A detailed literature study on

Table 3
Summary of the Explicit and Implicit Teaching Environment in the Lesson-Unit "Explaining the First World War"

Learning phase	General design principles	Teacher and learning activities *explicit* condition	Teacher and learning activities *implicit* condition
Preparatory phase(s) Students *connect* the topic to the broader domain and their prior knowledge	Raising situational interest	Teacher fosters a sense of rooted relevance by presenting a funny nonhistorical analogy intended to trigger causal reasoning (Chapman, 2003) Students read, listen, generate ideas, answer	Teacher fosters a sense of rooted relevance by showing a short video-clip about the murder on Franz Ferdinand and asking whether one murder could be responsible for a war to start? Students watch, generate ideas, answer
Instructional phase(s) Students *understand* the task and the goals of the task	Working on (an) open (sub)task(s) Social interaction Making teacher & student thinking visible	Teacher uses the nonhistorical analogy to explicate and model thinking about concepts and strategies connected to historical causation Students read, listen, generate ideas, answer	Teacher uses the video clip to connect students' first-order knowledge with the key question; introduces task and goals Students watch, listen, generate ideas, answer
Phase(s) of application Students *construct* a causal historical explanation		Teacher coaches, scaffolds, provides feedback (focus on strategies and second-order concepts) Students work in triads to coconstruct an explanation Students engage in card-sorting tasks and concept-mapping to select, categorize and connect causes Students develop a vocabulary for categorizing and connecting causes by applying a wordlist (Woodcock, 2005) Students write a mini-essay (supported by the wordlist)	Teacher coaches, scaffolds, provides feedback (focus on first-order knowledge) Students work in triads to coconstruct an explanation Students engage in card-sorting task to select causes and organize causes in a graphical organizer Students synthesize their analysis in a written synopsis that connects causes and answers the question Students prepare a presentation to communicate their analysis
Phase(s) of integration Students *discuss and compare* their explanations		Teacher asks (epistemological) questions, addresses misconceptions, provides feedback; related to causal strategies and concepts Students present concept-maps and essays, and engage in whole-class discussion Students compare and reflect on similarities and differences in products, the concepts used and the strategies engaged Students reflect on epistemological questions	Teacher asks questions, addresses misconceptions, provides feedback; related to first-order knowledge Students deliver presentations and engage in whole-class discussion Students compare and reflect on similarities and differences of products (focus on first-order knowledge) Students exchange perspectives and arguments

the principles in both lesson-units can be found in Stoel et al. (2015).

The explicit teaching environment. The major difference between the lesson-units was the explicit attention to strategies, second-order concepts, and epistemological questions related to historical causation. This attention was operationalized for each phase of the learning environment and consisted of teacher-led activities (instruction and scaffolding), student-led activities (group work), and shared activities (whole-class reflection).

In the instructional phases, the teacher explicated relevant second-order concepts and modeled the targeted strategies connected to historical causation. A nonhistorical analogical story was used to start students' thinking about multicausality, causal categories, and connected second-order concepts and to develop a multilayered model of causal relationships. Furthermore, the teacher modeled and discussed different ways to verbalize causal explanations with various degree of causal "nuance."

In the application phases, students practiced the relevant strategies and concepts by working together on causal (sub)tasks. Card-sorting tasks (involving events, people, countries, developments, and phenomena in Europe prior to World War I) and concept maps were used to stimulate analysis of different (types of) causes, to draw causal connections, and to reflect on appropriate causal models and the roles these causes played in their explanations. Students were equipped with a wordlist to scaffold their verbalization of causal connections (Woodcock, 2005). In the final lesson, the groups constructed a miniessay based on their analysis. In this phase, the teacher's role was mainly to scaffold and coach.

In the phases of integration, whole-class discussion was deployed to verbalize, broaden, and deepen students' understanding. Guided by the teacher, students reflected on their categorization of causes, their causal model and connections (as witnessed in their concept maps), and on their miniessays. Furthermore, whole-class discussion was used to reflect on epistemological questions about the constructed nature of students' explanations, the differences between interpretations, and the criteria for assessing the quality of these explanations.

The implicit teaching environment. Students in the implicit condition worked on the same task as did students in the explicit condition. Also in this condition, raising situational interest and collaborative learning were important design principles. In the lesson unit, we differentiated between the same four instructional phases. The crucial difference between lesson-units was that triads in the implicit condition worked on the whole task and constructed a causal explanation without paying explicit attention to causal strategies, concepts, and epistemological questions. The learning activities in this lesson-unit were designed to balance the amount of analysis and synthesis of first-order knowledge in the explicit condition alongside the alternation of writing, discussing, and visualizing, and the total time students interacted with the historical content. Instruction and constructive feedback in all phases focused on first-order knowledge and on supporting task execution.

In the instructional phase, the teacher activated prior knowledge by showing a short video clip about the murder of Archduke Franz Ferdinand to introduce the key question and, subsequently, to explore students' initial ideas. This phase focused on explicating the first-order knowledge that had been the subject of the two preparatory lessons. Afterward, the open-ended task and the goals of the task were introduced to the students.

In the phases of application, a card-sorting task (involving the same cards as in the explicit condition) and a graphical organizer (an empty presentation format) were used to select and organize events, developments, and phenomena connected to the outbreak of World War I. Based on this scheme, the triads subsequently wrote a synthesis text in which they linked their causes together and answered the key question. This synopsis formed the backbone of the PowerPoint presentation, which students designed in the second lesson, and of their subsequent oral presentation.

The final lesson aimed at integrating students' knowledge; the groups presented their causal explanations, compared each other's work, and gave feedback. This activity aimed to broaden and deepen students' (first-order) knowledge. Guided by the teacher, students reflected on each other's explanations, focusing primarily on first-order knowledge. Implicitly, of course, the learning activities confronted the students with several different interpretations and with causal concepts and connections. However, these issues were not explicitly addressed.

Research Instruments

Because our theoretical framework conceptualized causal-historical reasoning to be underpinned by different types of knowledge (knowledge of second-order concepts and strategies, and epistemological beliefs), we not only measured students' ability to construct a causal-historical explanation in a pre- and posttest but also assessed underlying knowledge and beliefs. At the posttest, students received two open prompts asking them to reflect on their learning gains and to provide a heuristic for future causal analysis in history.

Because the MDL considers first-order knowledge to be an important element in the development of expertise, we measured historical-topic knowledge both before and after the experiment. The MDL also conceptualizes a learning environment—focusing on epistemological questions, deep-level strategies, and connected concepts—to stimulate the development of individual interest. This too was measured both before and after the experiment; however, we did not expect individual interest to increase in only three lessons.

Finally, students' situational interest was measured at the end of the third intervention lesson as a control variable. Based on the MDL, we wanted to make sure that (a) both conditions were successful in arousing students' situational interest and (b) would do so to a comparable extend to control for potential differences in motivational quality of both conditions (because this might confound attributing effects to differences in cognitive approach). In our analyses, situational interest was used as a covariate.

Knowledge of causal reasoning strategies and second-order concepts. We administered a 19-item questionnaire twice (as a pretest and a posttest) to measure students' knowledge of second-order concepts and causal-reasoning strategies. Students had to score items on a six-point Likert scale, ranging from 0 (*strongly disagree*) to 6 (*strongly agree*). The questionnaire was based on literature and expert consultation, and a previous version was used in the pilot study (Stoel et al., 2015). Reliability analysis led to the exclusion of four items that lowered scale reliability. Fifteen items yielded a Cronbach's alpha of .64 (n = 82) at the pretest and .68

(n = 89) at the posttest. These items were used in the analysis. The following are examples of items: "in an historical explanation it is important to differentiate between different roles causes might have played"; "an historical explanation is usually constructed as a chain of causes and consequences" [recoded]; and "in an historical explanation you must also explain how causes interact."

Epistemological beliefs. Students' epistemological beliefs were measured twice (in a pretest and a posttest). We used a translated version of the Beliefs About History Questionnaire (BHQ) developed by Maggioni (2010). To explore the translated questionnaire, two datasets were collected prior to the experiment—one in another research project (N = 140) and one in the pilot study (N = 74). In both datasets, the copier scale yielded unacceptable low reliability. Therefore, we decided to exclude this scale from the current questionnaire. Furthermore, two criterialist items were excluded from the questionnaire because the translated versions did not load with the other criterialist items in both datasets. This could be the result of shifted meaning due to translation, or it could be connected to differences in historical culture between the United States and the Netherlands. The questionnaire used in this study, therefore, includes all subjectivist items from the original BHQ (9-items) and six out of eight items from the criterialist scale. In the results, we report students' (a) subjectivist epistemological beliefs and (b) their criterialist epistemological beliefs as separate dependent variables.

In the questionnaire students had to score items on a six-point Likert scale, ranging from 0 (strongly disagree) to 6 (strongly agree). We calculated scale reliability for the two remaining scales. The Cronbach's alpha for the subjectivist scale was .77 (n = 90) at pretest and .85 (n = 90) at posttest. Items in this scale related to the supposed subjective nature of historical knowledge. For example, "since there is no way to know what really happened in the past, students can believe whatever story they choose" and "good students know that history is basically a matter of opinion." The six items of the criterialist scale yielded a Cronbach's alpha of .65 (n = 88) at pretest and .65 (n = 92) at posttest. These items were all related to methodological criteria for constructing and judging historical interpretations. Example items included, "comparing sources and understanding author perspective are essential components of doing history" and "knowledge of the historical method is fundamental for historians and students alike."

Posttest open questions. During the posttest, students were also given two open prompts intended to explore their learning about causal-reasoning strategies, second-order concepts, epistemological reflections, and broader cognitive and motivational learning gains in a more reflective manner.

Heuristic prompt. This prompt asked students to provide a roadmap for when they would engage in future causal-historical inquiry (as an example, the collapse of the Soviet Union was mentioned). We developed a rubric consisting of one criterion (domain specificity) with two levels to code the heuristics. On the first level, students reported no heuristic, a fully generic one, or only very shallow references were made to causal concepts (i.e., only words such as *causes*, *consequences*, or *connections* were used). On the second level, students' answers at least centered on one dimension of causal reasoning (e.g., focusing on historical content, listing second-order concepts or referring to causal-reasoning strategies). Subsequently, we used the rubric to blindly code all the data. After explaining the rubric, the third author scored a subset of 31 random answers (29%). Interrater reliability was κ = .80.

Report of learning gains. This open prompt asked students to reflect on what they had primarily learned in the previous lessons. Three domain-specific codes were generated in accordance with the lesson goals: "looking for multiple causes" for answers that mentioned this basic causal-reasoning strategy, "drawing causal connections or using causal categories" for answers that referred to more sophisticated second-order concepts and causal-reasoning strategies, and "epistemological reflection" for answers focusing on the interpretative nature of causal analysis and the possibility of multiple valid answers. Furthermore, we defined one generic code ("additional") and gave this code to answers that, among others, referred to motivational aspects, the effectiveness of working with historical-inquiry tasks in general, historical content, and general study skills. Each category could only be scored once per student; although, an answer could be coded in multiple categories. An independent second rater scored a subset of 31 random answers (29%). Beforehand, the codebook was explained, the rater practiced on a subset and the differences were discussed. Interrater reliability for the four categories was multiple causes (κ = .81), causal connections and categories (κ = .77), epistemology (κ = .87), and additional (κ = .67).

Causal reasoning ability. Research has shown that reading multiple sources and writing argumentative accounts is an effective approach to elicit historical reasoning (Rouet, Britt, Mason, & Perfetti, 1996; van Drie, van Boxtel, & van der Linden, 2006; Wiley & Voss, 1999). Therefore, an explanatory-writing task was designed to measure students' ability to apply their first-order knowledge and their knowledge of causal concepts and reasoning strategies. The writing task was tested in a previous pilot study (Stoel et al., 2015). In the current study, students were asked to rewrite their pretest essay at the posttest. We expected that rewriting would heighten the sensibility of the instrument by lowering the complexity of the task at posttest. (Students were not required to read new sources or create a new written explanation but could revise their essay based on the newly gained knowledge.) Based on Rijlaarsdam, Couzijn, and van den Bergh (2004), who define rewriting as a goal-directed activity, we expected that this would allow students to more easily apply what they had learned in the intermittent lessons.

At pretest, students were asked to construct a 300-word explanation on why Germany became involved in World War I. Students were provided with a set of nine primary historical sources. Also, they were given a factsheet that listed the events, people, developments, countries, phenomena, and dates of the prewar period that had been discussed during the preparatory lessons. The factsheet was designed as a table to prevent a narrative (causal) template. The set of sources was constructed so that students could argue that Germany had provoked the war or had been "pulled in" by actions of other countries. Arguments of different types could be drawn from the sources (e.g., triggers, catalysts, and preconditions; direct and indirect causes; economic, political, and sociocultural causes; and personal agency).

A rubric consisting of four criteria was developed to analyze students' written work (see Appendix B). On each criterion, students received a score ranging between 0 and 2 points.

Students were scored on (a) the number of structural causes presented in their writing, (b) the number of structural causes substantiated by specific historical events (triggers or catalysts), (c) the explanatory model (linear, abstract list, abstract integrated), and (d) the use of nuanced second-order language and causal connections. All essays were blindly coded by two raters. For the first 31 essays, interrater correlation was not yet satisfactory; therefore, a final score for these essays was calculated based on agreement after discussion. For the remaining essays, interrater correlation was Pearson's $r = .71$ ($n = 156$). The mean score between the two raters was used for further analyses. Reliability of the four categories yielded a Cronbach's alpha of .70 ($n = 94$) at pretest and .66 ($n = 92$) at posttest.

First-order knowledge. A 17-item knowledge test was conducted twice (as a pretest and a posttest) to measure historical first-order knowledge. The test was slightly adapted from the test used in the pilot study (Stoel et al., 2015). The items were divided into four categories and measured students' (a) knowledge of prewar alliances (1 item), (b) ability to connect historical concepts to countries (12 items), (c) chronological knowledge (1 item), and (d) ability to generate concrete historical examples of abstract historical concepts (3 items). Two raters scored the three open items in all tests. Interrater reliability on these items was Pearson's $r = .85$ ($n = 182$). In the analysis, we used the scores of the second rater who had not been involved in the implementation of the experiment. A mean score was calculated for each category separately on a scale ranging from 0 to 1, and subsequently we calculated a pooled mean.

Individual interest. An 8-item questionnaire was conducted twice (in a pretest and a posttest) to measure students' individual interest in history. Students scored items on a six-point Likert scale, ranging from 0 (*strongly disagree*) to 6 (*strongly agree*). The individual interest questionnaire was based on an adaptation of the task-value scales from the Motivated Strategies for Learning Questionnaire (MSLQ) developed for mathematics education (Linnenbrink-Garcia et al., 2010; Pintrich, Smith, García, & McKeachie, 1993). Sample items included "I enjoy the school subject of history," "I can use historical knowledge well outside school," and "it is important for me to be able to think historically." Cronbach's alpha for the questionnaire was .89 ($n = 91$) for the pretest and .89 ($n = 93$) for the posttest.

Situational interest. A 12-item questionnaire was conducted at the end of the final intervention lesson to measure students' situational interest. Situational interest was measured to ascertain that differences in learning gains would not be attributable to differences in motivational quality of the lesson-units in both conditions. Furthermore, situational interest was used as a covariate in our model to prevent the confounding of learning outcomes with difference in interest in the learning environment. Students had to score items on a six-point Likert scale, ranging from 0 (*strongly disagree*) to 6 (*strongly agree*). The questionnaire was based on a validated questionnaire for mathematics education (Linnenbrink-Garcia et al., 2010). Sample items included, "what I have learned in these lessons is useful for me to know," "I liked what we learned in these lessons," and "these lessons were so exciting that I could easily maintain my attention." Cronbach's alpha for the questionnaire was .91 ($n = 91$).

Treatment Fidelity, Missing Data, Homogeneity of Triads, Statistical Procedure, Effect Sizes, and Homogeneity of Conditions

Treatment fidelity. The lesson-units were delivered by two external teachers who were both experienced teachers and historians with a firm grasp of the content knowledge, the strategies, and second-order concepts related to causal reasoning as well as the epistemological questions involved in historical explanations. For all lessons, detailed plans were designed—which established learning goals, teacher and student activities, and precise scripts for instruction and whole-class discussions. Both teachers engaged in a 2-day training session in which all lessons were meticulously discussed and prepared, leaving room for adaptation. During these meetings, a shared meaning and approach was negotiated and definitive lesson plans were determined. All lessons were discussed both before and after the execution. No important deviations from the plans were reported; the lessons could be executed as planned. To prevent difference between teachers and to avoid confounding learning outcomes, both external teachers switched conditions between classes—each person taught two explicit and two implicit subgroups. Furthermore, all student products were collected and compared on thoroughness. This comparison showed that students in both conditions were able to complete the tasks and that the quality of their work was satisfactory. Finally, students' engagement and interest in the lessons was measured by the situational interest questionnaire. On a six-point scale, situational interest was rated positive by students in both conditions ($M_{exp} = 4.08$, $SD = .69$, $n = 53$ and $M_{contr} = 3.91$, $SD = .64$, $n = 42$) and did not differ significantly between conditions.

Missing data. Missing value analysis showed that, on average 1.5% of the values were missing. Missing value count on the individual variables ranged between 0 and 5 ($M = 1.5$, $Mdn = 1$, mode = 1). The number of absentees varied between 0 and 3 at the different test moments ($M = 1.5$). Full information maximum likelihood approach (FIML; method = FCS; $n_{imputed} = 20$) was used in order to include all students in the analysis (Little, Jorgensen, Lang, & Moore, 2014).

Homogeneity of triads. All triads within the subgroups were composed of a high scoring, a low scoring, and an average student (based on students' history grades during the school year) to prevent the confounding of outcomes on dependent variables with differences between triads. After the experiment, all data was imported into R, and we ran a multilevel model to explore group effects on outcome variables. All of the intraclass correlations were 1% or less, showing that blocking on student achievement worked well and that the groups were indeed homogeneous. Therefore, there was no further need to run a multilevel analysis; a regular regression model would yield similar results.

Statistical procedure. In the analyses, univariate GLM's were used to analyze the mean scores on the six dependent variables (knowledge of causal concepts and strategy, subjectivist epistemological ideas, criterialist epistemological ideas, essay quality, first-order knowledge, and individual interest). The choice for univariate analysis was considered acceptable because of the conceptual distinctions between the measured constructs. This distinction is also reflected in the separate univariate hypotheses of our study. Moreover, the FIML approach for handling missing data does not allow for multivariate analysis. Pooled outcomes on the

imputed dataset ($n_{imputed} = 20$) can only be calculated as univariate regression. In the analysis, we controlled for students' pretest scores and for the differences in their situational interest.

Selection and calculation of effect size. Squared semipartial correlations were used as measures of effect size. In regression analysis, squared semipartial effect sizes yield the proportion of variability uniquely predicted by the independent variable when the other independent variables have been controlled (Fritz, Morris, & Richler, 2012). Semipartial square effect size is interpreted as follows: $rs^2 > .01$: small effect; $rs^2 > .09$: medium effect; $rs^2 > .25$: large effect (Cohen, 1988).

Homogeneity of conditions at pretest. To check the homogeneity of the conditions at pretest, a regression analysis was performed on all six dependent variables at pretest means, and we entered condition as a covariate. No significant univariate differences were found at the pretest (see Table 4). Subsequently, the mean scores on students' situational interest in the task were compared to check for potential motivational differences between the lesson-units in both conditions. Differences in motivational quality between conditions might confound the attribution of effects to a difference in cognitive approaches. As reported above, situational interest in both groups was positive and did not differ significantly between the conditions. The checks increased confidence in the homogeneity of both conditions at pretest and the comparability of the motivational qualities of both learning environments.

Results

Effects of Explicit Teaching on Students' Second-Order and Strategy Knowledge and Epistemological Beliefs

Knowledge of causal reasoning strategies and second-order concepts. Two exploratory paired-samples t tests on the nonimputed dataset showed that students' knowledge of causal strategies and second-order concepts increased significantly in the explicit condition, $t(51) = 4.20$, $p < .000$, but no significant change was found in the implicit condition. As we expected, students in the explicit condition scored significantly higher at the posttest than in the implicit condition, while controlling for differences in students' situational interest and pretest scores, $t(91) = 3.33$, $p = .001$, $sr^2 = .09$, post hoc power = .97.

Epistemological beliefs. Two exploratory paired-samples t tests on the nonimputed dataset showed that the level of agreement with subjectivist epistemological ideas increased significantly in the explicit condition, $t(52) = 2.20$, $p = .032$, but no significant change was found in the implicit condition. Students' subjectivist epistemological ideas, controlling for pretest scores and situational interest, differed significantly between conditions at posttest, $t(91) = 2.21$, $p = .027$, $sr^2 = .04$, post hoc power = .74. Students in the explicit condition reported a higher agreement with subjectivist beliefs at posttest compared with students in the implicit condition. This result contradicted the hypothesis that developing more nuanced ideas would lead students to become more critical toward subjectivist beliefs. This expectation was based on the theoretical framework of Maggioni et al. (2009).

Two exploratory paired-samples t tests on the nonimputed dataset showed that the value students attributed to criterialist epistemological ideas decreased significantly in the implicit condition, $t(39) = -2.08$, $p = .044$, whereas a nonsignificant increase was found in the explicit condition $t(52) = 1.85$, $p = .070$. Controlling for pretest scores and situational interest, a significant posttest difference was found for teaching condition on students' criterialist epistemological ideas, $t(91) = 2.60$, $p = .009$, $sr^2 = .04$, post hoc power = .87. Students in the explicit condition reported a higher value on items related to disciplinary criteria for generating historical knowledge compared with students in the implicit condition. This result was in line with our hypothesis.

Open prompts. The open prompts were analyzed for references to causal strategies and concepts and references to epistemological ideas. Because of the qualitative nature of the answers, no covariables were included in the analysis. Chi-square tests of measuring goodness of fit were performed to investigate differences between the explicit and implicit condition. Table 5 lists the descriptive statistics.

Table 4
Pretest and Posttest Means and Standard Deviations, Pooled Across Imputations, and Posttest Differences

Dependent variable	Dependent variable and condition	n	Pretest		Posttest		p^d	sr^{2d}
			M	SD	M	SD		
Knowledge of causal reasoning strategies[a]	Exp	53	4.16	.38	4.37	.41	.001	.09
	Contr	42	4.05	.32	4.08	.31		
Subjectivist beliefs (epistemology)[a]	Exp	53	3.22	.62	3.38	.69	.027	.04
	Contr	42	3.11	.63	3.03	.69		
Criterialist beliefs (epistemology)[a]	Exp	53	4.39	.49	4.51	.44	.009	.04
	Contr	42	4.41	.49	4.27	.50		
Essay[b]	Exp	53	.89	.39	1.24	.34		
	Contr	42	.76	.36	1.12	.39		
First-order knowledge[c]	Exp	53	.44	.17	.68	.15		
	Contr	42	.40	.17	.68	.13		
Individual interest[a]	Exp	53	3.79	.86	4.10	.67	.008	.02
	Contr	42	3.63	.80	3.72	.79		

Note. Exp = experimental condition; Contr = control condition.
[a] Min = 1, max = 6. [b] Min = 0, max = 2. [c] Min = 0, max = 1. [d] Posttest differences and effect sizes were calculated controlling for pretest scores and situational interest.

Heuristic prompt. The effect of explicit teaching on the domain specificity of the heuristic prompt was significant, $\chi^2(1, N = 95) = 13.20$, $p < .000$. In the explicit Condition 58% of the students referred to at least one dimension of causal-historical reasoning, compared with 21% of the students in the implicit condition.

Report of learning gains. Students in the explicit condition reported significantly more often to have learned about causal concepts or strategies, $\chi^2(1, N = 95) = 5.50$, $p = .019$. Examples of student answers in this category included, "it is important to draw relations between causes"; or "how you can classify historical facts, or events in a narrative, in triggers, catalysts, background causes." Furthermore, students in the explicit condition reported significantly more often to have acquired epistemological understandings, $\chi^2(1, N = 95) = 8.14$, $p = .004$. For instance, students report to have learned that "not everybody will consider the same causes as causes," that "history is mainly about explanation," or that "history can be viewed from multiple perspectives." In contrast, students in the implicit condition reported significantly more "off topic," general, and diffuse learning gains, $\chi^2(1, N = 95) = 18.63$, $p < .000$. In this category, answers were coded that referred to first-order content (e.g., "I remembered almost nothing about the First World War, but have learned a lot about it"), to motivational aspects (e.g., "in history, thorough research can quickly lead to forming an opinion. This is nice, because most people think differently about this"; or "it can become a more interesting and more fun subject if lessons are taught like this"), to the effectiveness of historical inquiry tasks (e.g., "doing historical research is far more complicated and complex than many people think"), or to general study skills (e.g., "you learn a lot by thorough reading and delving into it"). Finally, both groups reported approximately equally as often to have learned that history always involves multiple causes.

Causal Reasoning Ability

Two exploratory paired-samples t tests on the nonimputed dataset yielded a significant increase in essay quality in both conditions, $t_{exp}(50) = 7.23$, $p < .000$; $t_{imp}(40) = 6.77$, $p < .000$. Contrary to our hypothesis, however, controlling for situational interest and pretest scores no significant difference was found between the mean scores in both conditions at posttest, $t(91) = .57$, $p = .571$ (for descriptives, see Table 4). Subsequently, we explored the four underlying criteria of the rubric on which the essay score was based. Controlling for pretest scores and situational interest, a significant difference was found at posttest on one criterion of the rubric, namely, "use of second-order language and causal connectors." Students in the explicit group ($M_{pre} = .70$, $SD = .48$; $M_{post} = 1.30$, $SD = .54$) scored significantly higher than students in the implicit condition ($M_{pre} = .51$, $SD = .50$; $M_{post} = .96$, $SD = .44$), $t(91) = 2.59$, $p = .010$, $sr^2 = .06$, post hoc power = .95.

Table 5
Means, Standard Deviations of Open Prompts, and Differences Between Conditions

Prompt and code	Exp (n = 53) M	Exp (n = 53) SD	Contr (n = 42) M	Contr (n = 42) SD	p
Heuristic prompt	.58	.50	.21	.42	<.000
Learning gains					
Multiple causes	.51	.51	.38	.49	
Causal concepts and strategies	.47	.50	.24	.43	.019
Epistemology	.23	.42	.02	.15	.004
Additional, of which	.23	.42	.67	.48	<.000
first-order content	6%		21%		
motivational aspects	2%		7%		
effectiveness of open-ended tasks	4%		19%		
general study skills	8%		10%		

Note. Exp = experimental condition; Contr = control condition.

First-Order Knowledge

Two exploratory paired-samples t tests on the nonimputed dataset yielded a significant increase of first-order knowledge in both conditions, $t_{exp}(52) = 11.05$, $p < .000$; $t_{imp}(40) = 12.01$, $p < .000$. Regression analysis on the posttest first-order-knowledge scores, while entering students' pretest scores and situational interest as covariates, revealed no significant effect of teaching condition on first-order knowledge, $t(91) = -.64$, $p = .520$.

Individual Interest

Two exploratory paired-samples t tests on the nonimputed dataset showed that students' individual interest increased significantly in the explicit condition, $t(52) = 4.76$, $p < .000$, but we found no significant change in the implicit condition. Although we expected that individual interest would remain stable, the regression revealed a small significant effect of condition on individual interest when we controlled for pretest scores and situational interest, $t(91) = 2.67$, $p = .008$, $sr^2 = .02$, post hoc power = .87. The effect was tempered by controlling for situational interest because the two constructs correlated very strongly (Pearson's $r = .40$ at pretest; Pearson's $r = .60$ at posttest).

Relationships Between Different Constructs Underlying Causal Historical Reasoning

This study was designed on the premise that, besides first-order knowledge, causal historical reasoning is related to a student's conceptual second-order and strategy knowledge and epistemological ideas. Furthermore, a relationship is expected between the cognitive dimensions of historical reasoning and individual interest. Hereunder, the relationships between the dependent variables at different points of measurement (the pretest and the separate posttests for both conditions) will be presented. The correlation tables can be found in Appendix C.

The correlation tables showed a weak to moderate relationship between second-order and strategy knowledge and students criterialist epistemological beliefs both at the pretest (Pearson's $r = .26$, $p = .026$) as well as at the posttests (explicit, Pearson's $r = .35$, $p = .011$; implicit, Pearson's $r = .28$, *ns*). Students who attributed greater value to criteria for generating historical knowledge also scored higher on their knowledge of causal reasoning strategies and concepts.

At the posttest in the explicit condition, a moderate negative relationship was found between knowledge of causal concepts and

strategies, and students' subjectivist epistemological ideas (Pearson's $r = -.39$, $p = .004$). Students in the explicit condition that were more critical about regarding history as "just an opinion," scored higher on knowledge of causal-reasoning strategies and concepts.

The pretest showed that students' criterialist epistemological beliefs were moderately related to their individual interest (Pearson's $r = .39$, $p < .000$). At the posttest, the correlation between criterialist epistemological beliefs and individual interest increased in the explicit condition (Pearson's $r = .66$, $p < .000$), which can be considered a strong relationship. Students in the explicit condition, who were in greater agreement with criterialist epistemological ideas, also reported a higher individual interest in history. The strengthening of this relationship appeared to be caused mainly by the increased level of individual interest reported by the explicit condition at the posttest.

Essay quality at the pretest was found to correlate moderately with interest (Pearson's $r = .30$, $p = .003$). Students who reported a higher interest in history also tended to write better essays. At the posttest in the implicit condition, a strong relationship was found between students' essay scores and their first-order knowledge (Pearson's $r = .41$, $p = .006$) and between essay scores and knowledge of causal-reasoning strategies (Pearson's $r = .43$, $p = .009$). No significant relationships with essay quality were found at the posttest in the explicit condition.

The correlation tables showed a strong relationship between first-order knowledge and individual interest at the pretest (Pearson's $r = .43$, $p < .000$). This relationship was also found at the posttest in both conditions. Students who reported a higher value on individual interest also scored higher on the first-order-knowledge tests.

Discussion

The goal of this study was to investigate the effects of explicit teaching on developing students' ability to construct a causal historical explanation. Based on our theoretical framework, it was expected that learning to reason in a domain not only demands open-ended tasks, social interaction, and the stimulation of situational interest but also requires explicit, well-structured instruction and practice (i.e., explicit teaching environment). An important element in our theoretical framework was the "holistic" definition of explicit teaching. Our model asserted that explicit teaching should attend to students' knowledge of strategies and second-order concepts connected to historical causation as well as to epistemological questions involved in constructing historical explanations.

As we expected, knowledge of causal-reasoning strategies and second-order concepts improved in the explicit condition but not in the implicit condition. At the posttest the explicit condition scored significantly higher than students in the implicit condition. Therefore, we concluded that these strategies and concepts are learnable in an explicit teaching environment but that they do not develop spontaneously in the context of an inquiry task. Analysis of the learner reports did provide additional support for this conclusion. Of the students in the explicit condition, 47% of the students mentioned to have learned about "causal connections and categories" or explicitly mentioned second-order concepts, compared with 24% in the implicit group. The same difference in domain specificity was found in students' heuristics. In the explicit Condition 58% of the students made references to at least one dimension of causal-historical reasoning, versus 21% in the implicit group. Remarkably, students in both conditions mentioned approximately as often to have learned that causal analysis in history always involves looking at multiple causes. At least for 11th grade preuniversity students, this learning goal appears to be attainable even without explicit instruction.

Looking at the value students attributed to the epistemological scales (subjectivist beliefs and criterialist beliefs), a significant difference was found at posttest in students' subjectivist beliefs. Contrary to our expectations, students in the explicit condition rated the subjective nature of historical knowledge at posttest higher than students in the implicit condition. Maggioni et al. (2009), however, associated a higher score on the subjectivist scale with a more naïve position because these items present historical knowledge as an opinion. A possible explanation for this development could be that students in the explicit condition had just received three lessons that focused on the construction of causal interpretations and emphasized the interpretative nature of historical knowledge. Therefore, students' scores may primarily reflect a move away from more absolutist ideas about historical knowledge rather than a strengthening of the idea that historical knowledge is "mere personal opinion." This explanation is supported by the fact that students in the explicit condition valued the items belonging to the criterialist scale significantly higher than the students in the implicit condition. This finding was in line with our expectations. However, effect sizes for the differences between conditions regarding epistemological beliefs were small. The results might indicate that development in epistemological beliefs in history is more adequately conceptualized as a movement along two dimensions—certain or uncertain knowledge and weak or strong emphasis on disciplinary criteria—instead of in three distinct stances (cf. Schommer, 1993). More research is needed to explore this question.

Besides the (small) changes found in students' epistemological beliefs, analysis of the reports on learning gains revealed that 23% of the students in the explicit condition referred to epistemological aspects (e.g., reporting to have learned about history as an interpretation), compared with only 2% in the implicit condition. This difference constitutes a large effect. Based on these results, we conclude that in the explicit condition epistemological aspects shifted more into focus and constituted a tangible part of the learning environment even though epistemological beliefs did not strongly change in these three lessons.

Students' criterialist epistemological beliefs held a weak to moderate correlation with their knowledge of causal strategies and concepts at the pre- and the posttest. Students who attributed a higher importance to disciplinary criteria for constructing historical interpretations also scored higher on the knowledge of causal strategies and second-order concepts needed for these accounts. This finding is in line with the relationship between epistemological beliefs and strategic processing predicated by the MDL (Alexander, 2005).

Furthermore, our data revealed that criterialist epistemological beliefs correlated with students' individual interest in history. Students with a greater interest in history also reported a higher appreciation of the disciplinary criteria for constructing historical knowledge. At the pretest, a strong relationship was found, and it

even increased to Pearson's $r = .66$ (a very strong relationship) at the posttest in the explicit condition. This means that about 50% of the variance found in these two variables could be explained by their association. Although no causal inferences can be made, a possible explanation for this result might be that attention to epistemological questions may stimulate students' appreciation of history. Anecdotal support for this was found in the learner reports. One student, for example, stated to have learned that "history is mainly about explanation. Also, it can become a more interesting and more fun subject if lessons are taught like this." Future research should shed more light on the exact (causal) relationship between these constructs.

Students in the implicit condition also mentioned to have learned about conducting research (19%), although these answers were less domain specific and more diffuse—a result that was confirmed in their answers to the heuristic prompt. Judging from the self-reports and the level of situational interest, however, the open-ended task was valued by students in this condition as well. We found it interesting that without explicit attention to the disciplinary concepts, strategies and epistemological underpinnings, students did not appear to regard the inquiry as historical inquiry.

The quality of students' essays developed significantly in both conditions, but contrary to our expectations, no clear difference between conditions was found. Applying understanding of concepts and strategies in an explanatory rewriting task appears to be a difficult step for students. This result was found using an assessment rubric focusing on multiple criteria—by scoring multiple causes, substantiation of causes, text structure, and use of second-order language. When zooming in on a central aspect of the lesson-unit, "use of second-order language and causal connectors," a small but significant effect of explicit teaching was found at the posttest. Students in the explicit condition incorporated a richer vocabulary of causal connectors and concepts to differentiate between different types of causes in their essays. It may be that this aspect of causal-historical reasoning is more readily included in an explanatory text but that more profound changes (e.g., changes in the structure of the explanation) require explicit attention to additional demands (e.g., knowledge of the genre) and perhaps prolonged practice. This is supported by studies focusing on explicit teaching of strategies to analyze historical sources in which effects on students' writing were found (see De La Paz, 2005; De La Paz & Felton, 2010). These studies were longer and included strategies for writing historical essays. However, a study by van Drie, Braaksma, and van Boxtel (2015) found positive effects from a relatively short discipline-based writing instruction on student essays, focusing on historical significance. Future research may investigate the effects of explicit instruction of causal concepts, strategies, and epistemological questions in a longitudinal design and combine a focus on historical reasoning with explicit attention to the causal historical genres.

In line with our expectations, first-order knowledge increased significantly in both groups, without differences between the conditions. This indicates that teaching and engaging students in learning activities while focusing on causal reasoning strategies, use of second-order concepts, and epistemological reflection does not negatively influence students' learning about historical topics—even when considerable time and focus are invested in learning about causal skills. This finding is in line with findings from earlier studies (see Nokes et al., 2007; Reisman, 2012). It is possible that because students in the explicit condition engaged in more deep-level strategies, they more thoroughly processed the first-order knowledge. Because the pretest on first-order knowledge was administered before the two preparatory content lessons, first-order learning gains could also have resulted from these lessons.

A stable correlation in the data was found between first-order knowledge and individual interest. This finding is in line with the conceptualization of expertise in the MDL and provides an impetus to design learning-environments that stimulate epistemological reflection and "[aim] for rooted relevance" (Alexander, 2005). Although it had been 2 years since students had studied World War I, the relationship between first-order knowledge and interest was already found at the pretest. This relationship between prior knowledge and interest is in line with a review study on interest, prior knowledge, and learning (Tobias, 1994), as well as with the MDL.

Correlation analysis did not yield clear support for our model that students' performance on a causal writing or rewriting task is underpinned by students' first-order knowledge, knowledge of causal strategies and concepts, and epistemological beliefs. In the pretest, essay quality was only related to interest but in the posttest of the implicit condition, the relationship shifted to first-order knowledge and knowledge of causal strategies and concepts. In the explicit condition, no clear relationships for essay quality were found. A possible explanation for this may be that our rubric focused on specific aspects of causal-historical reasoning and, for instance, did not include use of first-order knowledge. Therefore, the relationship between first-order, second-order, and strategy knowledge, epistemological beliefs, and "deep historical analysis" (Nokes et al., 2007, p. 503) remains a point to be further explored. In the future, large scale research should further explore the relationships between interest, epistemological beliefs, knowledge (first-order, second-order, and strategy) and causal-historical reasoning skills. Structural equation modeling could be an important step in this direction.

Conclusion

Our study shows that, the explicit teaching of strategies and second-order concepts, within the context of an inquiry task, does constitute a prerequisite for developing students' conceptual and strategy knowledge connected to causal-historical reasoning. The causal reasoning questionnaire, the learner reports and heuristics, and students' essays provided evidence for how this knowledge was more effectively developed and shifted more into focus in the explicit-teaching condition. Based on individual interest and students' learner reports, this explicitness also appears to have contributed to the value that students ascribed to the learning environment. An important finding in this study was the importance of explicitly addressing epistemological beliefs in the history classroom. Although students' beliefs did not strongly change during the lessons, many students referred to the epistemological dimensions in their learner reports. A strong relationship was found between the value students attributed to academic criteria for assessing historical interpretations (criterialist epistemological beliefs) and their individual interest. This relationship increased at the posttest belonging to the explicit condition. These results suggest that

addressing these questions may stimulate individual interest. These findings are in line with the MDL (Alexander, 2005). We also found that applying the knowledge in an explanatory (re)writing-task remains a difficult step for students.

This study has several limitations. First, the study was designed to investigate effects of explicit teaching on student learning. We strove to reduce teacher effects by working with two external teachers, who took much time to internalize the pedagogical principles. Although this strengthened treatment validity, it may have reduced ecological validity. This leads to two follow-up questions: To what extent can teachers incorporate the principles in their practices? And what would be the learning effects of integrating such an approach into everyday practice?

Second, the experiment was only conducted among 11th grade preuniversity students and the sample consisted of 95 students from one school. These choices allowed for a tightly organized, randomized setup, but it also limits the generalizability of the results. Future research should explore the effectiveness of explicit teaching of causal strategies, concepts, and epistemological reflection on a wider variety of age groups, schools, and school levels.

Third, we strove to include instruments that had been validated in other studies (Linnenbrink-Garcia et al., 2010; Maggioni et al., 2004, 2009; Pintrich et al., 1993). However, due to the domain-specific nature of our questions, several instruments were designed within the context of this study—most notably the causal-reasoning strategies and concepts questionnaire, and the essay task. We designed these instruments based on a literature review, discussed them with experts in the field of history education, and tested them out in a pilot study. Future studies should provide more robust support for these instruments. This study underscored that writing and rewriting an explanatory essay is a complex task that not only demands causal-historical reasoning ability but also knowledge of the genre and general reading and writing skills. An interesting future study could develop instruments that allow us to measure causal-historical reasoning skills in a more direct manner—rather than through a writing task. Reisman (2012) developed such an instrument, but it focusses on historical reasoning about sources.

We believe that our study holds implications for practice. Based on our results, we suggest that open-ended tasks, social interaction, and students' sense of rooted relevance should be given a more prominent place in history education. These characteristics do not only provide a fruitful context for acquiring historical topic knowledge but also provide a starting point to develop students' historical (causal) reasoning skills. The lesson-unit provides evidence that students value the task, especially when combined with explicit teaching of concepts, strategies and epistemological questions. Such an approach appears to stimulate students viewing inquiry as historical inquiry and history as an interpretative subject. It allows students to acquire intended domain-specific, deep-level strategies and appears to stimulate individual interest—although more support may be needed to apply this knowledge in a writing task. Preservice and in-service training should support history teachers in learning to design and to implement explicit-teaching environments that foster historical reasoning. This support should focus on the content of this explicit teaching and on providing teachers with concrete learning activities and open-ended tasks.

References

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher, 32,* 10–14. http://dx.doi.org/10.3102/0013189X032008010

Alexander, P. A. (2005). Teaching towards expertise (Special Monograph on Pedagogy—Teaching for Learning). *British Journal of Educational Psychology Monograph Series II, 3,* 29–45.

Barton, K. C., & Levstik, L. (2004). *Teaching history for the common good.* Mahwah, NJ: Erlbaum.

Central Bureau for Statistics. (2015, December 8). *Inkomen van particuliere huishoudens met inkomen naar kenmerken en regio* [Data file]. Retrieved from http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=80594ned&D1=2,4&D2=1&D3=0&D4=0,121&D5=1&HDR=G2,G1,T&STB=G3,G4&VW=T

Chapman, A. (2003). Camels, diamonds and counterfactuals: A model for teaching causal reasoning. *Teaching History, 112,* 46–53.

Coffin, C. (2004). Learning to write history: The role of causality. *Written Communication, 21,* 261–289. http://dx.doi.org/10.1177/0741088304265476

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Routledge.

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator, 6,* 38–46.

De La Paz, S. (2005). Effects of historical reasoning instruction and writing strategy mastery in culturally and academically diverse middle school classrooms. *Journal of Educational Psychology, 97,* 139–156. http://dx.doi.org/10.1037/0022-0663.97.2.139

De La Paz, S., & Felton, M. K. (2010). Reading and writing from multiple source documents in history: Effects of strategy instruction with low to average high school writers. *Contemporary Educational Psychology, 35,* 174–192. http://dx.doi.org/10.1016/j.cedpsych.2010.03.001

Erdmann, E., & Hassberg, W. (Eds.). (2011). *Facing–mapping–bridging diversity: Foundation of a european discourse on history education* (Vol. 1). Schwalbach am Taunus, Germany: Wochenshau Verlag.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141,* 2–18. http://dx.doi.org/10.1037/a0024338

Halldén, O. (1997). Conceptual change and the learning of history. *International Journal of Educational Research, 27,* 201–210. http://dx.doi.org/10.1016/S0883-0355(97)89728-5

Havekes, H. G. F. (2015). *Knowing and doing history. Learning historical thinking in the classroom* (Doctoral dissertation). Retrieved from http://hdl.handle.net/2066/141364

Khishfe, R., & Abd-El-Khalick, F. (2002). Influence of explicit and reflective versus implicit inquiry-oriented instruction on sixth graders' views of nature of science. *Journal of Research in Science Teaching, 39,* 551–578. http://dx.doi.org/10.1002/tea.10036

King, P. M., & Kitchener, K. S. (2002). The reflective judgment model: Twenty years of research on epistemic cognition. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 37–61). Mahwah, NJ: Erlbaum.

Kuhn, D., & Weinstock, M. (2002). What is epistemological thinking and why does it matter? In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 121–144). Mahwah, NJ: Erlbaum.

Lee, P., & Shemilt, D. (2009). Is any explanation better than none? *Teaching History, 137,* 42–49.

Levstik, L., & Barton, K. C. (2008). *Researching history education: Theory, method, and context.* New York, NY: Routledge.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70,* 647–671. http://dx.doi.org/10.1177/0013164409355699

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. (2014). On the joys of missing data. *Journal of Pediatric Psychology, 39,* 151–162. http://dx.doi.org/10.1093/jpepsy/jst048

Maggioni, L. (2010). *Studying epistemic cognition in the history classroom: Cases of teaching and learning to think historically* (Doctoral dissertation). Retrieved from http://hdl.handle.net/1903/10797

Maggioni, L., Alexander, P., & VanSledright, B. (2004). At a crossroads? The development of epistemological beliefs and historical thinking. *European Journal of Social Psychology, 2,* 169–197.

Maggioni, L., VanSledright, B., & Alexander, P. A. (2009). Walking on the borders: A measure of epistemic cognition in history. *Journal of Experimental Education, 77,* 187–214. http://dx.doi.org/10.3200/JEXE.77.3.187-214

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development, 50,* 43–59. http://dx.doi.org/10.1007/BF02505024

Nokes, J. D., Dole, J. A., & Hacker, D. J. (2007). Teaching high school students to use heuristics while reading historical texts. *Journal of Educational Psychology, 99,* 492–504. http://dx.doi.org/10.1037/0022-0663.99.3.492

Pintrich, P. R., Smith, D. A. F., García, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement, 53,* 801–813. http://dx.doi.org/10.1177/0013164493053003024

Prangsma, M. E., van Boxtel, C. A. M., & Kanselaar, G. (2008). Developing a "big picture": Effects of collaborative construction of multimodal representations in history. *Instructional Science, 36,* 117–136. http://dx.doi.org/10.1007/s11251-007-9026-5

Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction, 30,* 86–112. http://dx.doi.org/10.1080/07370008.2011.634081

Rijlaarsdam, G. C. W., Couzijn, M., & van den Bergh, H. (2004). The study of revision as a writing process and as a learning-to-write process: Two prospective research agendas. In L. Allal, L. Chanqouy, & P. Largy (Eds.), *Revision: Cognitive and instructional processes* (Vol. 13, pp. 189–207). Boston, MA: Kluwer. http://dx.doi.org/10.1007/978-94-007-1048-1_12

Rouet, J.-F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88,* 478–493. http://dx.doi.org/10.1037/0022-0663.88.3.478

Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. In B. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 135–148). Englewood Cliffs, NJ: Educational Technology Publications.

Schommer, M. (1993). Epistemological development and academic performance among secondary students. *Journal of Educational Psychology, 85,* 406–411. http://dx.doi.org/10.1037/0022-0663.85.3.406

Seixas, D. P., & Morton, T. (2013). *The big six historical thinking concepts.* Scarborough, Ontario, Canada: Nelson College Indigenous.

Stoel, G. L., van Drie, J. P., & van Boxtel, C. A. M. (2015). Teaching towards historical expertise: Developing a pedagogy for fostering causal reasoning in history. *Journal of Curriculum Studies, 47,* 49–76. http://dx.doi.org/10.1080/00220272.2014.968212

Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research, 64,* 37–54. http://dx.doi.org/10.3102/00346543064001037

van Boxtel, C., & van Drie, J. (2013). Historical reasoning in the classroom: What does it look like and how can we enhance it? *Teaching History, 150,* 44–52.

van Drie, J., Braaksma, M., & van Boxtel, C. (2015). Writing in history: Effects of writing instruction on historical reasoning and text quality. *Journal of Writing Research, 7,* 123–156. http://dx.doi.org/10.17239/jowr-2015.07.01.06

van Drie, J., & van Boxtel, C. (2003). Developing conceptual understanding through talk and mapping. *Teaching History, 110,* 27–32.

van Drie, J., & van Boxtel, C. (2008). Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review, 20,* 87–110. http://dx.doi.org/10.1007/s10648-007-9056-1

van Drie, J., & van Boxtel, C. (2011). In essence I'm only reflecting: Teaching strategies for fostering historical reasoning through whole-class discussion. *International Journal of Historical Learning, Teaching, and Research, 10,* 55–66.

van Drie, J., van Boxtel, C., Jaspers, J., & Kanselaar, G. (2005). Effects of representational guidance on domain specific reasoning in CSCL. *Computers in Human Behavior, 21,* 575–602. http://dx.doi.org/10.1016/j.chb.2004.10.024

van Drie, J., van Boxtel, C., & van der Linden, J. L. (2006). Historical reasoning in a computer-supported collaborative learning environment. In A. M. O'Donnell, C. E. Hmelo-Silver, & G. Erkens (Eds.), *Collaborative learning, reasoning, and technology* (pp. 265–296). Mahwah, NJ: Erlbaum.

VanSledright, B. A. (2011). *The challenge of rethinking history education: On practices, theories, and policy.* New York, NY: Routledge.

VanSledright, B. A., & Limón, M. (2006). Learning and teaching social studies: A review of cognitive research in history and geography. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 545–570). Hillsdale, NJ: Erlbaum.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91,* 301–311. http://dx.doi.org/10.1037/0022-0663.91.2.301

Wineburg, S. S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past.* Philadelphia, PA: Temple University Press.

Woodcock, J. (2005). Does the linguistic release the conceptual? *Teaching History, 119,* 5–14.

Appendix A

Lesson Goals (Explicit and Implicit Condition)

Experimental phase	Goals
Preparatory topic lessons (1/2)	Students acquire knowledge of several concrete events, concepts, countries and dates in the period leading up to the First World War. Students acquire knowledge of several abstract phenomena (i.e., nationalism, imperialism, alliances, arms race).
Intervention lessons (3/4/5)	Both conditions; general Students improve their ability to construct causal historical explanations by engaging in causal analysis through an open-ended task that prompts them to select and organize possible causes and construct a causal explanation of the outbreak of the First World War. Explicit condition; explicit attention to: Students can explain that historical explanations always involve multiple causes. Students develop a vocabulary related to causal second-order concepts and causal connections, that is, (in)direct, short-term, economic, trigger, catalyst, precondition. When constructing a historical explanation, students can organize and classify causes within the dimensions of time, content and role. Students build a causal model in which causes exert both simultaneous and linear (direct or indirect) influences. Students can explain that causal explanations are never a copy of the past itself (copier stance). Multiple valid explanations can co-exist, but not all explanations are valid (subjectivist stance). There are (academic) criteria for assessing historical explanations, including the use of evidence and arguments (criterialist stance).

Appendix B

Rubric Explanatory Writing Task

Criterion	Beginning	Developing	Adequate	Points
Structural causes	The author mentions no or only one historically correct structural cause	The author mentions two historically correct structural causes	The author mentions three or more historically correct structural causes	Max 2
Substantiation of structural causes[a]	The author doesn't substantiate any structural cause with concrete historical event (incidental causes). OR: The author only superficially elaborates one or two structural causes (e.g., elaboration without using incidental causes, or without making clear the relationship between a structural and incidental cause)	The author substantiates one or two structural cause with concrete historical event (incidental causes). OR: The author superficially elaborates more than two structural causes (e.g., elaboration without using incidental causes, or without making clear the relationship between a structural and incidental cause)	The author substantiates more than two structural cause with concrete historical event (incidental causes)	Max 2
Explanatory model	Concrete. Author describes causality on a linear level.	Abstract. Author describes causality on an abstract level, but this genre is still in development. The structure of the text can be characterized as messy or incomplete	Abstract. The author describes causality on an abstract level and does so in an appropriate and structured manner	Max 2

(*Appendices continue*)

Appendix B (*continued*)

Criterion	Beginning	Developing	Adequate	Points
Use of second-order language/causal connections	Author uses no or little causal connectors of second-order language (this category also applies for students that only use "because" and "therefore," unless this is done in a very thorough manner)	Author uses causal connectors and second-order language, but almost completely aimed at organizing the text (i.e. first, multiple) AND/OR: Author makes adequate use of connection words (throughout the text, in a correct way, that makes clear the causal links)	Author uses causal connectors and second-order language, not only to organize but also to describe impact and directness (evaluate) AND/OR: Author uses a rich repertoire of causal connectors that describe nuanced relationships (and differences) between causes (i.e. this reinforced, in the background)	Max 2

[a] On this criterion argumentation is an important element (the relationship between structural and incidental causes must be described.

Appendix C

Correlations Between the Dependent Variables at Pretest and Posttest (Explicit and Implicit Condition)

Table C1
Correlations Between the Dependent Variables at the Pretest

Dependent variables	*n*	1	2	3	4	5	6
1. 2nd-order/strategy	95	—					
2. Subjectivist beliefs	95	.09	—				
3. Criterialist beliefs	95	.26*	−.01	—			
4. Essay quality	95	.08	−.12	.14	—		
5. 1st-order	95	.22*	.16	.10	.16	—	
6. Individual interest	95	.06	−.04	.39**	.30**	.43**	—

* $p < .05$ (2-tailed). ** $p < .01$ level (2-tailed).

Table C2
Correlations Between the Dependent Variables at Posttest (Explicit Condition)

Dependent variables	*n*	1	2	3	4	5	6
1. 2nd-order/strategy	53	—					
2. Subjectivist beliefs	53	−.39**	—				
3. Criterialist beliefs	53	.35*	−.01	—			
4. Essay quality	53	.15	−.10	.23	—		
5. 1st-order	53	.07	.04	.14	.24	—	
6. Individual interest	53	.14	.14	.66**	.11	.30*	—

* $p < .05$ (2-tailed). ** $p < .01$ level (2-tailed).

(*Appendices continue*)

Table C3
Correlations Between the Dependent Variables at Posttest (Implicit Condition)

Dependent variables	*n*	1	2	3	4	5	6
1. 2nd-order/strategy	42	—					
2. Subjectivist beliefs	42	.07	—				
3. Criterialist beliefs	42	.28	.05	—			
4. Essay quality	42	.43**	−.11	.12	—		
5. 1st-order	42	.02	−.19	−.19	.41**	—	
6. Individual interest	42	.15	−.16	.26	.22	.49**	—

* $p < .05$ (2-tailed). ** $p < .01$ level (2-tailed).

Received August 3, 2015
Revision received April 29, 2016
Accepted May 17, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 3, 338–354

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000147

Process Mediates Structure: The Relation Between Preschool Teacher Education and Preschool Teachers' Knowledge

Sigrid Blömeke
Centre for Educational Measurement at the University of Oslo

Lars Jenßen and Marianne Grassmann
Humboldt University of Berlin

Simone Dunekacke
Leibniz Institute for Science and Mathematics Education

Hartmut Wedekind
Alice Salomon University of Applied Science

Data about processes and outcomes of preschool teacher education is scarce. This paper examines the opportunities to learn (OTL) of prospective preschool teachers (N = 1,851) at different types and stages of preschool teacher education and their relation to general pedagogical knowledge (GPK), mathematics pedagogical content knowledge (MPCK), and mathematical content knowledge (MCK) with standardized tests. Process indicators in terms of OTL and outcome indicators in terms of knowledge varied substantially across teacher education types and stages. Controlling for preschool teachers' background, multilevel models revealed that OTL in general pedagogy and mathematics pedagogy provided during teacher education were significantly related to GPK and MPCK. Effect sizes reached up to 2 thirds of a standard deviation. OTL in mathematics pedagogy were in turn significantly related to the type of institution that offered a program in favor of pedagogical colleges compared with vocational schools. OTL were also significantly related to program stage in favor of the last year of preschool teacher education compared with the beginning. Process characteristics in terms of OTL mediated fully or partly structural characteristics of teacher education such as type of institution or program stage. These results suggest that the OTL provided are more important than whether prospective preschool teachers were at the beginning or the end of their program or whether they were prepared at vocational schools or pedagogical colleges (although entrance differences have still be taken into account). It may be an important responsibility of policymakers then to ensure that all prospective preschool teachers receive sufficient OTL.

Keywords: early childhood education, opportunity to learn, pedagogical knowledge, mathematics, preschool teachers, teacher education

This article shows that opportunities to learn general pedagogy and mathematics pedagogy by prospective preschool teachers during their teacher education program were related to their general pedagogical and mathematics pedagogical content knowledge. How many opportunities to learn prospective teachers received was in turn often related to the type and stage of a teacher education program in favor of pedagogical colleges when compared with vocational schools and in favor of the last year of preschool teacher education when compared with the beginning of their education. These findings provide important information to understanding how preschool teachers gain their professional knowledge, and these results can assist policymakers in deciding how to improve preschool teacher education. The results suggest that in particular opportunities to learn mathematics pedagogy provided during preschool teacher education may be more important for knowledge acquisition than more distal factors such as the type of institution where prospective teachers are prepared. This in turn suggest that it may be worthwhile to focus reforms of preschool teacher education more directly on opportunities to learn instead of on less direct structural changes.

Highlights

- Standardized tests of prospective preschool teachers' knowledge were developed.
- Objectivity, reliability and content, construct and criterion validity was confirmed.
- Domain-specific opportunities to learn were strongly related to GPK and MPCK.
- Type of teacher education institution and program stage were related to OTL.
- Relation of teacher education structure to outcomes was partly mediated through processes.

This article was published Online First September 1, 2016.

Sigrid Blömeke, Centre for Educational Measurement at the University of Oslo; Lars Jenßen and Marianne Grassmann, Department of Education, Humboldt University of Berlin; Simone Dunekacke, Department of Mathematics Education, Leibniz Institute for Science and Mathematics Education; Hartmut Wedekind, Department of Science Education, Alice Salomon University of Applied Science.

Lars Jenßen is now at Freie Universität Berlin.

The research project "Structure, level and development of prospective preschool teachers' competencies in mathematics" (KomMa) was funded by the German Federal Ministry of Education and Research (BMBF) as part of the research initiative "Modeling and measuring competencies in higher education (KoKoHs)" (FKZ 01PK11002A).

Correspondence concerning this article should be addressed to Sigrid Blömeke, Centre for Educational Measurement at the University of Oslo (CEMO), N–0318 Oslo, Norway. E-mail: sigribl@cemo.uio.no

Research on effects of preschool teacher education on prospective preschool teachers' knowledge and skills has so far mostly been restricted to distal indicators of teacher knowledge such as degrees or licensing (Whitebook, Gomby, Bellm, Sakai, & Kipris, 2009). The same applies to studies examining the opportunities to learn (OTL) of prospective preschool teachers, in the present paper defined as content coverage providing the chance to gain the knowledge and skills necessary to succeed with fostering the development of preschool-age children (i.e., 3 to 6 years of age). In most studies, OTL were operationalized through distal indicators such as the length or the type of a preschool teacher education program (Bogard, Traylor, & Takanishi, 2008). Corresponding to the state of research on primary and secondary school teacher education (Abell Foundation, 2001; Darling-Hammond, 2000), results from research on preschool teacher education have been contradictory. Whereas some studies have established significant relations between preschool teacher education and preschool teachers' knowledge and skills or long-term outcomes such as children's development (Burchinal, Cryer, Clifford, & Howes, 2002; Howes, Whitebook, & Phillips, 1992; Tout, Zaslow, & Berry, 2005; Whitebook et al., 2009), other studies have failed to establish relations (Early et al., 2007).

Most authors agree that this unsatisfactory state of research is attributable to problems with the measures that have been used. Degrees and licenses but also program length and types are rather imprecise (i.e., unreliable) indicators of the knowledge and skills that preschool teachers gain during their education or the OTL they encounter during teacher education. The meanings of the distal indicators depend on the specific norms and practices applied in different teacher education institutions (Carroll, 1963). Educational effectiveness research has revealed that the content that is covered (Berliner, 1985) and the time allocated to such OTL (Carroll, 1963) are at the core of teaching and learning (Travers & Westbury, 1989). Standardized instruments for assessing preschool teacher educations' OTL and their outcomes in such a specific way are missing but urgently needed (Bogard et al., 2008; Early et al., 2007).

Furthermore, there is no systematic framework that is able to define the structure of preschool teachers' knowledge and skills and conceptualize their dimensions in more detail. Nor is there a systematic framework beyond institution-specific curricula that conceptualizes the OTL offered during preschool teacher education. Such frameworks are therefore urgently needed as well. The distal indicators currently in use (e.g., degree or program length) are only rough approximations. They are not indicative of specific *domains* such as reading or mathematics, let alone sufficiently specific with respect to details within these domains.

State of Research

A summary of the state of research on the relation between preschool teacher education and teacher knowledge reveals substantial holes. Furthermore, because of the lack of a shared understanding of the construct "OTL" and the lack of standardized measures, different authors operationalize OTL differently which leads to some ambiguity in the following review as well.

A nationally representative U.S. study found that domain-specific OTL in mathematics, reading, or science are scarce during preschool teacher education because even at institutions of higher education, most programs focus on general pedagogical OTL (Early & Winton, 2001; see also Isenberg, 2000). Linguistic and cultural diversity or the education of children with disabilities were additional blind spots (Lobman, Ryan, & McLaughlin, 2005). In an analysis of preschool teachers' self-reports, another U.S. study correspondingly found that they did not feel sufficiently prepared to teach children with diverse backgrounds (Ryan, Ackerman, & Song, 2004).

If one takes into account research on professional development (PD) after initial teacher education, a clearer picture emerges. Hamre et al. (2012) found substantial effect sizes in the relation between OTL offered to preschool teachers and their ability to perceive the classroom accurately as assessed with a standardized video test as well as between OTL and the teachers' ability to support children's literacy skills as assessed with a standardized test. Similar results were found by Pianta et al. (2014). However, Piasta, Logan, Pelatti, Capps, and Petrill (2015) were able to provide evidence for PD effects on performance in preschools only in the domain of children's science learning but not in mathematics learning.

Thus, we have only initial evidence that OTL matter in that they are related to preschool teachers' knowledge and skills. Despite frequent pleas for more research on the specific effects of preschool teacher education on teacher characteristics with standardized and domain-specific measures (Early et al., 2007; Whitebook et al., 2009), not many studies have undertaken this effort though. Whereas recently a large number of studies using direct, standardized, and domain-specific teacher assessments has been published on primary and secondary teacher education—confirming strong links between OTL during teacher education and teacher education outcomes in terms of prospective teachers' knowledge, which in turn predicted teaching performance and student achievement (Blömeke, Suhl, Kaiser, & Döhrmann, 2012; Tatto et al., 2012; Voss, Kunter, & Baumert, 2011)—preschool teacher education is still a "black box."

One particular blank spot exists with respect to the effectiveness of preschool teacher education below the tertiary level which applies to many developing countries but also to a range of Southern and Western European countries (Wallet, 2006). Such programs do not take place at institutions of higher education but at postsecondary or secondary vocational school. Completion of high school is thus not necessarily a requirement for entering a preschool teacher education program.

In many countries, policy efforts have been undertaken to move preschool teacher education up to the tertiary level. A prominent example of this is the Head Start program in the U.S., which is directed toward providing high-quality preschools to low-income children. The program receives funding only on the condition that half of its preschool teachers hold a bachelor's degree (Bassok, 2012). Graduates from Bachelor programs have thus become the main target population of preschool research (see, e.g., Early et al., 2007). However, with a few exceptions such as some Scandinavian countries and some states in the U.S., preschool teachers with a degree below the tertiary level are still the majority in the U.S. (Bogard et al., 2008) and in many other countries, including Germany, which is the context of the present study.

To overcome the research gaps described above, the objective of this paper is to unpack the black box of "preschool teacher education" by examining the relation between domain-specific OTL

provided during preschool teacher education and domain-specific outcomes in terms of preschool teachers' knowledge while controlling for their background characteristics. All knowledge dimensions were assessed in a standardized way in a multicohort, multigroup design to be able to include prospective preschool teachers from different types of teacher education institutions. The instruments were developed on the basis of a conceptual framework derived from educational effectiveness research, which will be presented in this paper as well. We paid particular attention to differences between institutions of higher education that award a bachelor's degree to preschool teachers and vocational schools part of the secondary school level.

How important it is to clarify the relation between preschool teacher education and teacher knowledge is demonstrated in studies that focused on the relation between this knowledge and the cognitive development of children. Early et al. (2006) found that preschool teachers with a bachelor's degree delivered higher mathematics-related instructional quality as indicated by standardized on-site observations and achieved stronger outcomes in a direct assessment of children's mathematical literacy than preschool teachers without such a degree. Preschool teachers' knowledge in mathematics, assessed directly with a standardized test, also significantly predicted their ability to perceive preschool situations appropriately and to perform instructional activities that support the development of children's mathematics literacy as assessed with a standardized video test (Dunekacke, Jenßen, & Blömeke, 2015a). Evidence exists in other domains (e.g., reading literacy) as well (Connor, Morrison, & Slominski, 2006; Landry, Anthony, Swank, & Monseque-Bailey, 2009).

Preschool Teacher Education in the Context of Germany

Preschool education in Germany is voluntary and can be subdivided into institutions covering 1- to 3-year-olds and institutions covering 3- to 6-year-olds. Teachers of the latter represent the target population of this study. At this age, more than 90% of the children are enrolled at least part-time—mostly in morning sessions—although parents have to pay a small fee (Statistisches Bundesamt, 2014). Preschools are typically run by local municipalities, churches (mostly Protestant or Catholic), or charity organizations, and some are organized privately with a special pedagogical profile. Preschools are not part of the school system but of the child and youth welfare system. They are therefore assigned to ministries of family affairs instead of ministries of education in the 16 German states so that there is more emphasis on care than on formal education.

Play-based activities represent the norm for teacher-child interactions (Liegle, 2008). Preschools organize these activities either in fixed groups with one full-time preschool teacher (or equivalent part-time employees) assigned to about 10 children or in larger groups of variable sizes looked after by teams of preschool teachers. Because more and more evidence points to the relevance of child development before schooling for later student achievement (see, e.g., Duncan et al., 2007), the belief that it is important to foster young children's cognitive development has increased in recent years—in particular with respect to 3- to 6-year-olds. All 16 German states have recently implemented standards for preschools that present ambitious cognitive objectives with respect to early reading, mathematics, and science literacy. This means that teachers have to use the informal context of preschool more often than before to foster these abilities. However, a systematic accountability system to support the achievement of these ambitious objectives does not yet exist.

Because of society's increased awareness of the relevance of preschool education, parents have recently earned the right to send their children to preschool beginning at age 1 when the paid parenthood leave ends. If a municipality is not able to offer such a child a spot in a preschool, parents are reimbursed for the private daycare costs that exceed the small fee they would have to pay for a spot in a regular preschool.

Preschool teachers are trained differently in the 16 German states. Typically, a two-tiered system exists. The majority of preschool teachers (more than 90%) are trained at vocational schools that provide teacher education on the secondary or post-secondary level. This means that the entrance requirement is not completion of high school but of 9 or 10 years of general schooling followed by 2 to 4 years of vocational training in a care profession (or a similar type of education). In parallel, there are also pedagogical colleges that are part of the higher education system and award a bachelor's degree. Fifty-six colleges existed at the time of our study in 2015. Students must have completed high school followed by a 6 to 12-month pedagogical internship to enter these colleges. Currently, only about 5% of preschool teachers have undergone this type of education, and the numbers are growing only slowly (Statistisches Bundesamt, 2014). The 16 German states are responsible for the preschool teacher education curricula; the 56 pedagogical colleges have the academic freedom to design their curricula so that the training conditions vary substantially across Germany.

Conceptual Framework

To the best of our knowledge, there is no conceptual framework that specifically describes the structure of preschool teachers' knowledge. To avoid a purely operational definition, we therefore applied basic educational-psychological dimensions of primary teachers' knowledge to preschool teachers but operationalized these on the basis of research on 3 to 6-year-old children's development and learning. This approach ensured connectivity between subsequent educational stages (Anders, 2012) so that we could examine the specifics of each one.

Preschool Teachers' Knowledge

According to Shulman (1986) and Weinert (2001), teacher knowledge is a multidimensional construct that includes general pedagogical knowledge, pedagogical content knowledge, and content knowledge. With respect to content, the present study was restricted to the domain of mathematics learning. Preschool teachers' knowledge then includes mathematics content knowledge (MCK), pedagogical content knowledge of how to foster mathematics learning in children between the ages of 3 and 6 (MPCK), and general pedagogical knowledge of how to organize the informal learning environment of preschool in general (GPK).

To define these dimensions in more detail, we conducted two systematic analyses of all preschool teacher education curricula from the 56 pedagogical colleges and the 16 federal states (for the vocational schools) as well as of all preschool standards imple-

mented in the 16 federal states (Jenßen, Dunekacke, & Blömeke, 2015). Preschool standards set by the 16 German states were used to describe the objectives of preschool with respect to children's mathematical learning. Preschool teacher education curricula were used to describe the OTL in mathematics, mathematics pedagogy, and general pedagogy offered to prospective preschool teachers at the different institutions in the 16 states. Construct maps (Wilson, 2005) summarized the results of the systematic analyses of preschool teacher education curricula and standards in terms of subdimensions and specific descriptors. These were used to represent the range of pedagogical and mathematical OTL and preschool objectives. During test development, these descriptors were operationalized with items that were represented in the majority of standards and curricula and were also supported by the literature (for detailed results, see the Appendix).

GPK includes general foundations from educational theory, psychology, and instructional research related to early childhood and learning processes of 3 to 6-year-olds (Blömeke, Jenßen, Dunekacke, Suhl, Grassmann, & Wedekind, 2015). An OECD (2004) review of early childhood curricula in five countries revealed that the present framework is in alignment with discussions elsewhere. *MPCK* includes diagnosing children's developmental state in mathematics and designing an informal learning environment that fosters the mathematical learning of children between the ages of 3 and 6 (Dunekacke, Jenßen, & Blömeke, 2015b). Again, this framework resembles discussions in other countries (NAEYC, 2009). *MCK* includes numbers, sets, and operations; shape, space, and change; quantity, measurement, and relations; data, combinatorics, and chance (Dunekacke et al., 2015a). Although developed in the national context of Germany, this framework also reflects discussions that are taking place elsewhere (Clements, Sarama, & DiBiase, 2004; National Research Council, 2009).

To ensure that the tests also included different cognitive processes, a second framework was developed on the basis of cognitive psychology (Anderson & Krathwohl, 2001). On the one hand, the items had to assess the recalling, understanding, and applying of knowledge as well as the knowledge-based generation of strategies. On the other hand, they had to capture cognitive complexity in terms of the different numbers of cognitive steps necessary to solve an item as well as different types of problem representations (Embretson & Daniel, 2008). The two frameworks, the alignment of frameworks and measures, as well as the inferences to be drawn from these measures, have been validated in a range of studies (Blömeke et al., 2015; Dunekacke et al., 2015a, 2015b; Jenßen, Dunekacke, Eid, & Blömeke, 2015).

OTL Provided During Preschool Teacher Education

Characteristics of preschool teacher education that potentially have an effect on prospective preschool teachers' GPK, MPCK, and MCK because of the differences in OTL provided, are the *type* of institution where a program takes place (in the present study: pedagogical college vs. vocational school) and the program *stage* (the beginning vs. end of a program). In samples of prospective primary and secondary teachers, there is evidence that these aspects of German teacher education matter in favor of longer programs—typically also requiring stronger entrance characteristics—on the one hand, and in favor of students at the end compared with students at the beginning of teacher education (Blömeke, Kaiser, & Lehmann, 2008; Kleickmann et al., 2013). The sample of the present study will therefore be drawn according to these structural characteristics of preschool teacher education, and we will test corresponding hypotheses (see below H2a, b).

However, such structural characteristics of institutions are proxies rather than direct measures of the teaching and learning processes going on. According to educational effectiveness research, OTL in terms of the content that was covered has to be taken into account (Berliner, 1985; Carroll, 1963). OTL reflect preschool teachers' chances to acquire GPK, MPCK, and MCK. OTL item development followed the same conceptual framework as applied for the three knowledge tests (see Appendix).

With respect to primary and secondary teacher education, evidence exists that such domain-specific proximal measures of teaching and learning processes are significantly related to outcomes (Blömeke, Suhl, & Kaiser, 2011; Blömeke et al., 2012; König, Blömeke, Paine, Schmidt, & Hsieh, 2011). GPK was significantly related to OTL in general pedagogy, whereas OTL in mathematics were significantly related to MCK and MPCK. OTL in mathematics pedagogy were significantly related to MPCK only when MCK was not included. These results applied both to primary and to secondary teachers. The aim of the present study is to expand this state of research to prospective preschool teachers by testing corresponding hypotheses (see below H1a, b, c, d).

Results from educational effectiveness research also revealed that in addition to examining such direct OTL effects on outcomes, indirect effects also need to be examined—for example, whether distal predictors such as structural characteristics of teacher education are mediated by proximal process indicators (see with respect to primary teachers Scheerens & Blömeke, 2016). Such a hypothesis is applied to preschool teacher education in this study as well (see below H3).

Hypotheses

H1: We hypothesized significant positive relations between different domain-specific process indicators of teaching and learning during preschool teacher education and corresponding teacher education outcomes. More specifically, we hypothesized that OTL in general pedagogy would have a stronger positive relation to GPK than OTL in mathematics pedagogy or in mathematics would (H1a). At the same time, we hypothesized that OTL in mathematics pedagogy would have a stronger positive relation to MPCK than OTL in general pedagogy or in mathematics would (H1b). Finally, we hypothesized that OTL in mathematics would have a stronger positive relation to MCK than OTL in general pedagogy or in mathematics pedagogy would (H1c).

Furthermore, we hypothesized that OTL would predict knowledge in a similar way in all subpopulations, which means technically that the relations would be invariant across prospective preschool teachers at pedagogical colleges and vocational school as well as across students at the beginning and at the end of teacher education (H1d).

H2: We hypothesized that structural characteristics of preschool teacher education would significantly positively predict prospective preschool teachers' knowledge. More precisely, prospective preschool teachers from pedagogical colleges were hypothesized

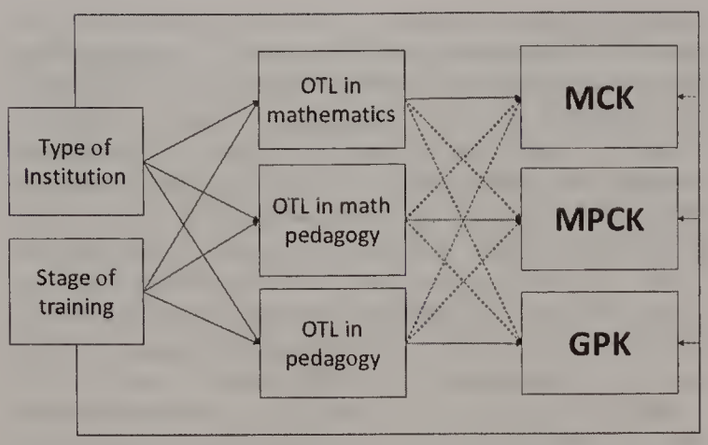


Figure 1. Hypothesized research model (OTL = opportunities to learn; MCK = mathematics content knowledge; MPCK = mathematics pedagogical content knowledge; GPK = general pedagogical knowledge; dotted lines were hypothesized to be weaker than solid lines).

to have significantly higher MCK, MPCK, and GPK compared with students from vocational schools (H2a). In addition, we hypothesized that MCK, MPCK, and GPK would be higher at the end compared with the beginning of preschool teacher education in both institutions, indicating progress during a program (H2b).

H3: Finally, we tested a mediation model (see Figure 1). We hypothesized that the relations of structural teacher education characteristics—type of institution and program stage—to MCK, MPCK, and GPK would be at least partly if not fully mediated through process characteristics in terms of the respective domain-specific OTL.

Method

Participants

The sample included 1,851 prospective preschool teachers from 86 classes in 44 teacher education institutions. Each class had between 6 and 82 students (M = 21). The 44 institutions included 31 of 516 vocational schools in Germany with a total of 67 classes ($M_{Stud/Class}$ = 20, SD = 7.9, Range = 6 to 46) and 13 of 56 German pedagogical colleges with a total of 19 classes ($M_{Stud/Class}$ = 25, SD = 18.8, Range = 6 to 82). From most institutions one class participated in the study but from a few vocational schools up to four classes participated.

The sample was drawn via personal contacts as a first step and by randomly contacting vocational schools and pedagogical colleges as a second step. Because preschool teacher education is the responsibility of the states in Germany, in this second step, care was taken to include all 16 states and to represent the larger states by including a larger number of schools from them than from the smaller states. Only a few institutions that we contacted were not willing to participate. Any institution that declined was replaced by another randomly drawn institution from the same state.

Because the relations of OTL to outcomes on the one hand and the differences between preschool teacher education at the secondary and higher education levels on the other hand were important research foci, we included four groups that were tested at the same time (see Table 1): prospective preschool teachers at the end and at the beginning of teacher education at institutions offering secondary education (vocational schools) and higher education (pedagogical colleges). Prospective preschool teachers in higher education were purposefully oversampled because otherwise the group would have been too small for scaling purposes given that it is a small minority of all prospective preschool teachers (about 5% only; Statistisches Bundesamt, 2014).

The consent of test takers was obtained by pointing out before the assessment started that participation in the study was voluntary and that beginning to fill out the forms was taken as consent. Those who did not want to participate were given the opportunity to leave the room at that moment. The instructions included the additional information that every participant could leave the room at any time and that consent could be withdrawn at any time until the tests were collected. None of the participants used this option, but it is possible that a small number of students did not come to school on the day of testing because they may have heard about it beforehand. Table 2 provides an overview of the sample's major characteristics.

The descriptive statistics were in line with our expectations and the demographics of the target population. The teachers in our sample who were at the end of preschool teacher education were 2 (vocational schools) or 4 (pedagogical colleges) years older than those who were at the beginning. Female teachers represented the majority in all four subgroups, and teachers' language background was almost always German. The biggest differences between the subsamples existed with respect to the two indicators of prior knowledge (school degree and number of years of mathematics in school) and the two indicators of socioeconomic background (mother's education and number of books at home). On each of the four indicators, the participants from vocational schools were at a disadvantage compared with the higher education students. The latter group reflects the average of the corresponding age group in the German population overall with respect to mother's education (Statistisches Bundesamt, 2010, p. 26).

Measures

Preschool teacher education outcomes: GPK, MPCK, and MCK. Test development was applied according to the conceptual framework described (for details, see the Appendix). A large item pool (n = 117) was developed in a joint effort between academic and practical experts from preschool mathematics, mathematics pedagogy, and general pedagogy. Item selection was ap-

Table 1
Sample Size

Type of institution	Vocational school (secondary education)	Pedagogical college (higher education)	Overall
Program stage			
First year	594 (32%)	287 (15%)	881 (47%)
Last year	774 (42%)	196 (11%)	970 (53%)
Overall	1,368 (74%)	483 (26%)	1,851 (100%)

Table 2
Descriptive Statistics of the Sample by Subgroup

Variable	First year vocational school	Last year vocational school	First year pedagogical college	Last year pedagogical college
Mean age in years (range)	22 (17–53)	24 (18–54)	22 (18–47)	26 (19–53)
Gender (female)	85%	83%	90%	90%
German language background (always spoken at home)	88%	89%	83%	86%
No. of books at home (>200)	23%	24%	41%	44%
Mother's education (at least a high-school degree)	17%	16%	32%	27%
Participant's own education (at least a high-school degree)	36%	44%	99%	99%
No. of years of mathematics in school (≤10)	47%	48%	2%	5%

plied on the basis of a series of cognitive labs and unstandardized prepilot studies as well as on the basis of standardized pilot (n = 454 prospective preschool teachers) and validation studies (n = 354 prospective preschool teachers; for results, see Dunekacke et al., 2015a, 2015b; Jenßen, Dunekacke, Eid, et al., 2015) as well as on the basis of conceptual considerations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). The three resulting knowledge tests consisted of multiple-choice, bundled, and open-response items. In all cases, gender-neutral language was used to reduce the risk of stereotype threats (Cadinu, Maass, Rosabianca, & Kiesner, 2005) and the language level was kept relatively simple to reduce bias that would favor students at pedagogical colleges.

The assessment of prospective preschool teachers' *MCK* consisted of 24 items that covered the four subdimensions numbers, sets, and operations; shape, space, and change; quantity, measurement, and relations; data, combinatorics, and chance as confirmed by expert validation (Dunekacke et al., 2015a). Open responses (including drawing figures and finishing tables or formulas) were required for 14 items, whereas 10 were multiple-choice items. These data resulted in 24 dichotomous items that were used to create the MCK score.[1] Score reliability was estimated according to Raykov, Dimitrov, and Asparouhov (2010) and was good (Pγ = .88). Figure 2 presents an example item (for scientific purposes, access to the full instrument can be granted by the first author of this paper; all items were administered in German, those displayed in the following were translated into English for the purposes of this publication).

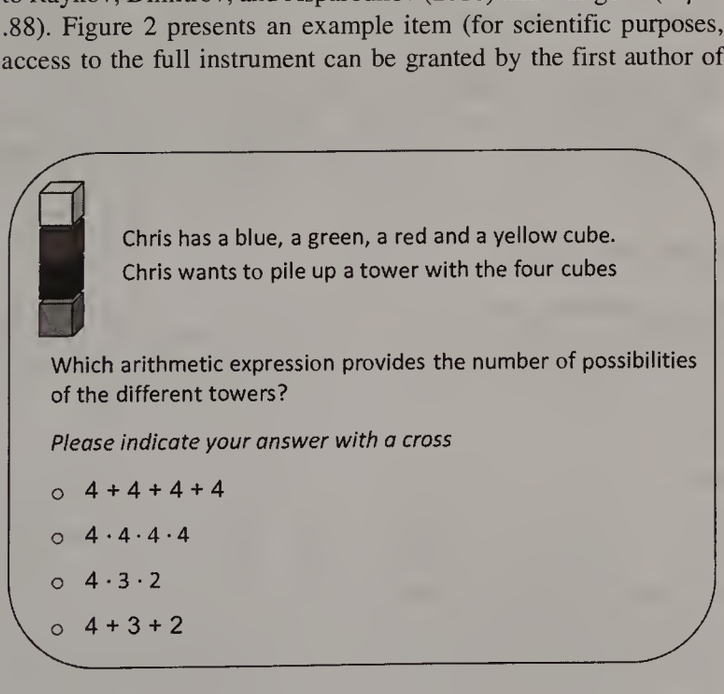


Figure 2. Example item from the MCK test (translated).

The *MPCK* assessment consisted of 28 items that covered diagnosing children's developmental state in mathematics and designing an informal learning environment that fosters the mathematical learning of children between the ages of 3 and 6 (Dunekacke et al., 2015b). Open responses were required by five items, whereas 23 were multiple-choice or bundled items. All items were scored dichotomously right or wrong so that the resulting MPCK score consisted of 28 items. Score reliability was good (Pγ = .87). Figure 3 presents an example item.

The assessment of *GPK* consisted of 18 items that covered general foundations from educational theory, psychology, and instructional research (Blömeke et al., 2015). Open responses were required by three items, whereas 15 were multiple-choice or bundled items. The information from these items was used to create 18 dichotomous items. Score reliability was lower than for the other two knowledge constructs but still sufficient (Pγ = .68). An example item is displayed in Figure 4.

Psychometric properties of the knowledge tests. To ensure sufficient objectivity in the implementation of the assessments, all procedures such as the timing or use of materials were prescribed in a manual, and administrators of the assessments were trained in a standardized way according to it. Evaluation objectivity was ensured by developing a codebook that described precisely how to code open-ended answers according to their content and which codes to evaluate as correct (1) or incorrect (0). Interrater reliability was ensured by coding 20% of the open-response items twice, resulting in a good interrater reliability for GPK (Md_{Kappa} = .76, Range = .64 to .88; Md_{Yules} = .98, Range = .95 to 1.00), MPCK (Md_{Kappa} = .73, Range = .64 to .92; Md_{Yules} = .97, Range = .92 to 1.00), and MCK (Md_{Kappa} = .78, Range = .69 to .86; Md_{Yules} = .99, Range = .95 to 1.00; Cohen, 1960; Yule, 1912).

The content validity of the three knowledge tests was confirmed in a standardized procedure by an expert panel. The experts evaluated each single item as well as the entire tests on their representativeness of the respective constructs and their power to predict and explain differences in response behavior of prospective preschool teachers (Jenßen, Dunekacke, & Blömeke, 2015).

[1] Separate subscores for each subdimension were not estimated because the distinction between these served merely conceptual purposes in the context of test development.

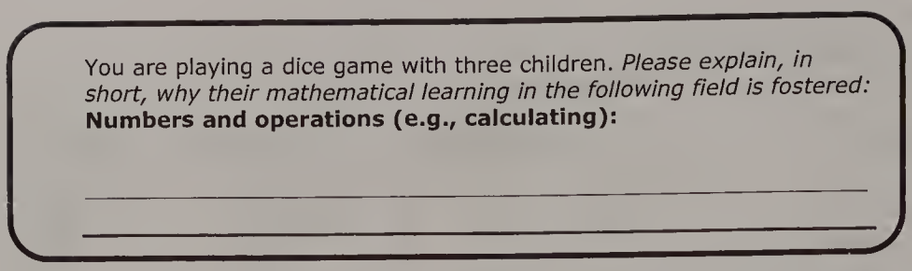


Figure 3. Example item from the MPCK test (subdomains data and modeling; translated).

Factorial validity of inferences drawn from the knowledge test results was confirmed with different samples from the pilot and validation studies by comparing the fit of a three- and a one-dimensional model to the data. The data revealed a better fit of the three-dimensional model. Although all three knowledge dimensions were significantly positively related with each other as hypothesized, it was still possible—again as hypothesized—to distinguish them empirically (Jenßen, Dunekacke, Baack, et al., 2015).

With the large present sample, it was in addition possible to carry out multiple-group confirmatory factor analysis (MG-CFA; Jöreskog, 1971) which means that each dimensional model was estimated in parallel within the four subsamples—prospective preschool teachers at the end and at the beginning of teacher education and this at secondary education and at higher education institutions thus also accounting for the oversampling of the latter. Future teachers represented the first level, and classes represented the second level in these models (two-level CFA mixture modeling using the known-class and cluster options implemented in MPlus 7.3; Muthén & Muthén, 2014).

Results supported the notion of preschool teachers' knowledge as a three-dimensional construct with latent correlations varying from .62 to .92. As hypothesized, the relation between GPK and MCK was the lowest, whereas the strongest relation existed between GPK and MPCK. However, given this strong relation a more parsimonious solution was estimated for exploratory reasons with two dimensions that unified the two latter. This model revealed a similarly good model fit as the three-dimensional model, which indicates that it may not be possible to distinguish GPK and MPCK empirically (see Table 3; Blömeke et al., 2015). Because such a two-dimensional model of preschool teacher knowledge has not yet been replicated with an independent sample and because the OTL provided during teacher education revealed a clear-cut three-dimensional structure (see below), the present study was carried out by applying such three-dimensional models.

Convergent and discriminant validity of the inferences drawn from the knowledge test results were supported in relation to school marks based on data from the present sample. The better a prospective preschool teacher was in mathematics at school, the higher the teacher's MCK and MPCK scores were ($\beta = .21$ or $\beta = .11$, respectively). By contrast, no significant relation existed between school mathematics and GPK (Blömeke et al., 2015).

Metric measurement invariance of the three knowledge tests was confirmed across the four different subgroups of the present sample of prospective preschool teacher education students at the beginning or the end of their program at vocational schools or pedagogical colleges as well as across students of different genders or with different language backgrounds (see Table 4; Blömeke et al., 2015). This means that it was possible to compare relations between constructs across these groups.

The criterion validity of inferences drawn from the test results in terms of their relation to performance was supported by data from an earlier sample. MCK and MPCK were direct predictors of prospective preschool teachers' abilities to perceive teacher-children interactions in preschool ($\beta_{direct} = .45$ or $\beta_{direct} = .60$, respectively) and were indirect predictors, mediated through perception skills, of their abilities to plan actions that foster children's mathematical development ($\beta_{indirect} = .43$ or $\beta_{indirect} = .58$, respectively) demonstrated in a video-based assessment that showed typical preschool situations (Dunekacke et al., 2015a, 2015b).

As hypothesized on the basis of Ma (1999), an application of latent-state-trait models to an earlier sample revealed that the

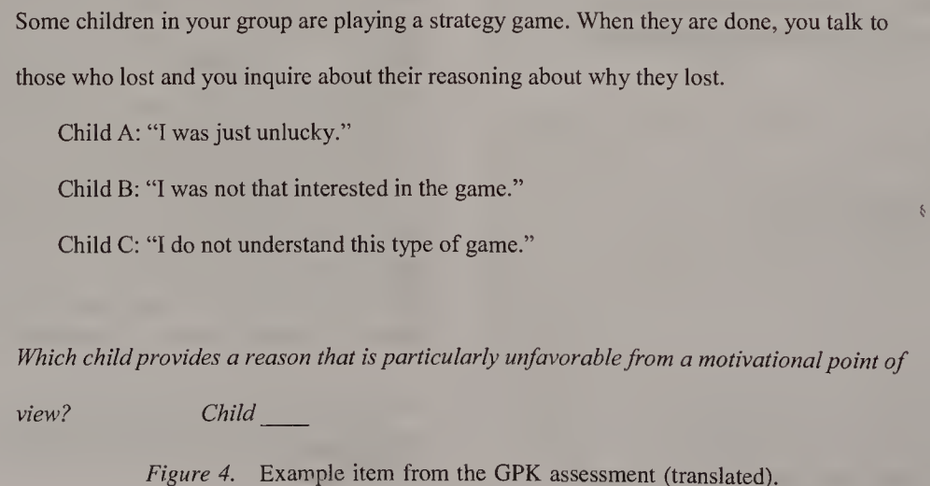


Figure 4. Example item from the GPK assessment (translated).

Table 3
Fit of One-, Two-, and Three-Dimensional Models of Prospective Teachers' Knowledge

Modell	# Par	LL	SCF	AIC	BIC_{adj}
One-dimensional	143	−42 535.4	1.92	85 356.9	85 692.4
Three-dimensional	146	−42 398.1	1.90	85 088.3	85 430.8
Two-dimensional	144	−42 403.3	1.90	85 094.7	85 432.6

Note. # Par = No. of parameters; LL = log likelihood; SCF = Scaling Correction Factor; AIC = Akaike's Information Criterion; BIC_{adj} = adjusted Bayesian Information Criterion.

different subdimensions of MCK and mathematics anxiety were negatively related ($\Psi = -.24$; $\Psi = -.38$; Jenßen, Dunekacke, Eid et al., 2015). Furthermore, MCK turned out to be stable enough over the course of 3 weeks to be regarded as a trait rather than a state.

Classroom-level predictors. OTL are modeled on the class level because the teacher education institutions are in control of the content and materials they deliver to the prospective preschool teachers. These are assigned to classes that take largely the same OTL. Neither vocational schools nor pedagogical colleges offer substantial possibilities to choose between content topics.

The prospective preschool teachers reported their OTL in mathematics, mathematics pedagogy, and general pedagogy by rating the coverage of certain topics in each field on 4-point Likert scales (1 = *not at all*, 4 = *intensely*). The topics were derived from the conceptual framework described in detail in the Appendix. OTL scores represent average item scores so that 1.0 represents the lowest score possible, 4.0 the highest score, and 2.5 the neutral point. Scale reliability for the present sample was evaluated with Cronbach's alpha, and model fit was evaluated with absolute and relative goodness-of-fit statistics derived from a CFA of the three constructs: OTL in mathematics, mathematics pedagogy, and general pedagogy (Hu & Bentler, 1999). Comparative Fit Index (CFI) estimates $> .95$ indicate a very good fit, and estimates $> .90$ a good model fit. Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) estimates $< .05$ indicate a very good fit, and estimates $< .08$ a good model fit.

OTL in mathematics were assessed with four items that covered numbers, sets, and operations; shape, space, and change; quantity, measurement, and relations; as well as data, combinatorics, and chance. The scale score's reliability and its model fit were good, $\alpha = .83$; $\chi^2(2) = 2.85$, $p = .24$; CFI = 1.00; RMSEA = .02, 90% CI [.00, .05], $p = .94$; SRMR = .01. OTL in mathematics pedagogy were surveyed with seven items that covered the extent to which the prospective preschool teachers had learned to diagnose the mathematical development of children such as their understanding of numbers, shapes, or measurement and to design informal learning environments that foster children's mathematical development in everyday situations or play. The reliability was very good and the model fit was satisfactory, $\alpha = .92$; $\chi^2(19) = 151.00$, $p < .001$; CFI = 1.00; RMSEA = .06, 90% CI [.05, .07], $p = .02$; SRMR = .04. Finally, the prospective preschool teachers reported their OTL in general pedagogy with four items that covered foundational topics such as basic terms of education and care, teaching methods, or dealing with heterogeneity. The reliability was just satisfactory ($\alpha = .75$), but the model fit was very good, $\chi^2(5) = 12.79$, $p = .03$; CFI = 1.00; RMSEA = .03, 90% CI [.01, .05], $p = .96$; SRMR = .05.

On the basis of an MG-CFA (see Table 5), factorial validity with respect to the hypothesized multidimensional structure of OTL was supported by the data. As indicated by the chi-square difference test and the difference in the information criteria reported, the three-dimensional model fit the data significantly better than the one- and the two-dimensional ones. In addition, the factor loadings of all but one item were above the critical threshold of .50 suggested in the literature (Dawis, 2000; Wheaton et al., 1977), whereas this applied to only half of the items in the one-dimensional and to 11 of the items in the two-dimensional models. Finally, the underlying factor significantly explained variance in the preschool teachers' response behavior of all items in the two- and the three-dimensional models, whereas this did not apply to four items in the one-dimensional model.

Control variables. All hypotheses were tested by controlling for the individual (level 1) background characteristics of prospec-

Table 4
Testing of Measurement Invariance of the Three Knowledge Tests Across Different Subgroups

	No. parameters			Log likelihood			Model comparison						*p* value		
							χ^2			*df*					
Model	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Configural	575	287	287	−41 894.0	−40 485.8	−40 528.0	—	—	—	—	—	—	—	—	—
Metric	374	220	220	−41 999.2	−40 514.9	−40 567.8	200.7	50.5	48.2	201	67	67	ns	ns	ns
Scalar	173	153	153	−42 216.5	−40 584.4	−40 617.1	381.2	138.1	97.9	201	67	67	<.05	<.05	<.05

Note. *df* = degrees of freedom; ns = non significant. Subgroups: Model 1 = 4 groups (beginning or end of teacher education at vocational schools or colleges of education), Model 2 = 2 groups (male, female teachers), Model 3 = two groups (language always German, not always German); model comparisons: metric vs. configural, scalar vs. metric based on the Satorra-Bentler scaled chi-square difference test TRd implemented in Mplus using the MLR chi-square and the scaling correction factor (Bryant & Satorra, 2012).

Table 5
Multiple-Group Confirmatory Factor Analysis

OTL model	Log likelihood	No. of parameters	AIC	BIC	BIC_{adj}	# items a)	# items b)
One-dimensional	−33,595.5	66	67,322.9	67,687.1	67,477.4	7	4
Two-dimensional	−32,471.6	71	65,085.3	65,477.1	65,251.5	11	0
Three-dimensional	−31,854.7	78	63,865.5	64,295.9	64,048.1	14	0

Note. OTL = opportunities to learn; AIC = Akaike information criterion; BIC = Bayesian information criterion; BIC_{adj} = sample-size adjusted Bayesian information criterion; # items a) = # of items with standardized factor loadings ≥ .50; # items b) = # of items where the variance was not significantly explained by the underlying latent variable. Chi-square difference test: $\chi^2(5)_{1\_2} = 1{,}123.9$, $p < .001$; $\chi^2(7)_{2\_3} = 616.9$, $p < .001$; $\chi^2(12)_{1\_3} = 1{,}740.8$, $p < .001$.

tive preschool teachers typically found to be predictive of educational outcomes such as gender, family background, and prior knowledge (Blömeke et al., 2012; Klusmann, Kunter, Voss, & Baumert, 2012; Rowan, Correnti, & Miller, 2002; Teddlie & Reynolds, 2000). Gender was coded dichotomously (0 = *female*, 1 = *male*). The language spoken at home as a first indicator of family background was assessed by using a 4-point Likert scale ranging from *never* speaking German at home through *always*. We dichotomized this scale for further analyses into *always* (1) versus the other categories (0) to make up for the skewed distribution (see Table 2). Mother's education as a second indicator of family background was measured on a scale that represented eight educational levels (*below lower secondary* through *PhD*). Because of the skewed distribution, it was also dichotomized into a category that included at least a high-school degree (1) versus the lower categories (0). Prior knowledge was assessed with self-reports of participants' most recent marks (class grades based on exams) in school mathematics and German as indicators, which have been shown to be valid indicators of prior knowledge in earlier studies (Blömeke et al., 2012). All background characteristics were introduced simultaneously on level 1 into the models.

Study Design and Scaling

Participants had 90 min to work on the instruments during the teacher education class they were enrolled in at the time the study was carried out. The instruments were presented in a paper-and-pencil format. We used six test booklets that were randomly distributed in each classroom. Each booklet started with background and OTL variables before the knowledge items followed in a multimatrix design. The items from each knowledge dimension (i.e., MCK, MPCK, and GPK) were distributed across five blocks (A1, A2, B, C, D) in such a way that each dimension was represented in each block, and the item difficulty of each block was the same on average according to the item difficulty estimates from the pilot studies. The blocks were distributed across the six test booklets so that A1 and A2 were represented in each booklet, whereas B, C, and D were randomly assigned. The blocks were then rotated in such a way that their sequence varied systematically.

To scale the knowledge test data, we applied the so-called Birnbaum model, a 2-parameter logistic item response theory (IRT) model that estimates not only item difficulties but also item discrimination parameters (Andrich, 2004).[2] The common items from Blocks A1 and A2 served as anchor items. The four subgroups from our sample were defined to have equal weights in the scaling process so that the characteristics of one group could not dominate the results. Missing values attributable to the booklet design of the tests could be regarded as missing completely at random and therefore did not introduce bias (Rubin, 1976). Missing values on items that were skipped or not reached were coded as missing at random (Pohl, Gräfe, & Rose, 2014). The proportion of the two latter types of missing values was low on all measures with a range of less than 1% to slightly above 4%. These missing values were included in the model estimation in a model-based iterative process by applying the full-information-maximum-likelihood (FIML) method, which uses all information available and is least prone to bias (Lüdtke, Robitzsch, Trautwein, & Köller, 2007). The resulting person estimates were transformed into a mean of 50 and a standard deviation of 10 to facilitate interpretation.

Data Analysis

The data were gathered in a multilevel structure with prospective preschool teachers (individual level, Level 1) nested in teacher education classes (classroom level, Level 2). Therefore, two-level structural equation modeling (SEM) was applied to test all hypotheses except H1d which was tested in a two-level confirmatory factor analysis (CFA) where the fit of models with fixed versus freely estimated covariances between OTL and knowledge in the four subgroups was compared. The intraclass correlations of .18 (GPK), .19 (MPCK), and .15 (MCK) indicated larger homogeneity within classrooms than if the prospective preschool teachers had been drawn randomly, thus justifying the multilevel approach. Explicitly modeling the cluster structure offers several advantages. First, statistically efficient estimates of regression coefficients and correct standard errors are obtained (Hox, 2002). Second, and this was important in the context of this paper, covariates at any level of the hierarchy can be used, and this makes it possible to examine the extent to which differences in achievement could be predicted by OTL or structural characteristics of preschool teacher education as class-level variables while controlling for individual preschool teachers' background. The individual-level variables were introduced by grand-mean centering (Snijders & Bosker, 2012).

In a few teacher education institutions up to four classes took part in the study which means that there may be some shared variance at the third (school) level. However, sample size was not

[2] We refrained from applying a 3-parameter model that would also have included a guessing parameter because sample size was not sufficient to do so and would have resulted in the risk of unstable model estimation.

sufficient to take this third level into account because variance on the second level would not have been sufficient with most institutions participating with only one class.

Because the relations between predictors and outcomes might vary across subpopulations, a multiple-group structure was added to the two-level models, which means that all path coefficients were allowed to vary across all groups when testing the hypotheses. *Within* each subpopulation, it could reasonably be assumed that the relations played out in the same way, which means that slopes were fixed across classrooms. We applied a robust maximum-likelihood estimator that could take into account the nonindependence of observations due to cluster sampling and would result in the estimation of robust standard errors (Muthén & Muthén, 2014). In all multilevel models, the EAP estimates obtained from the 2PL IRT scaling of MCK, MPCK, and GPK were used as manifest indicators to make these models less complex and identifiable.

In light of the sample size and the moderate complexity of most models, the 1% level of significance was used (with the exception of the more complex mediation model that tested H3 for which the 5% level was used). We report Cohen's *d* (1988) as the measure of effect sizes with respect to mean differences in predictors and outcomes between the four subgroups examined in this paper. Estimates larger than $d = 0.2$ can be regarded as small, larger than $d = 0.5$ as medium, and larger than $d = 0.8$ as large effects. Differences in regression coefficients were tested for statistical significance based on Clogg, Petkova, and Haritou (1995).

Results

Descriptive Statistics

OTL in mathematics and mathematics pedagogy were offered to a lower degree than OTL in general pedagogy, and these findings held at both institutions and at the beginning as well at the end of teacher education (see Table 6). The differences in OTL reported between the beginning and the end of teacher education had very large effect sizes as indicated by Cohen's *d*, and this held in all three domains. Prospective preschool teachers at vocational schools reported more OTL than their counterparts at pedagogical colleges at the beginning of their programs, particularly in general pedagogy. By contrast, prospective preschool teachers from pedagogical colleges reported an advantage in OTL in mathematics pedagogy at the end of their programs with a very large effect size.

GPK and MPCK differed significantly between vocational schools and pedagogical colleges in favor of the latter (see Table 7), and this finding held at the beginning as well as at the end of teacher education. Effect sizes indicating the differences were larger at the end ($d = 0.74$ or $d = 0.88$, respectively) than at the beginning of the programs ($d = 0.31$, $d = 0.50$). Both groups of students had significantly more GPK and MPCK at the end of their teacher education, but again, these differences were larger at pedagogical colleges (around $d = 0.62$) than at vocational schools ($d = 0.21$ or $d = 0.32$, respectively). MCK also differed substantially between vocational schools and pedagogical colleges in favor of the latter (around $d = 0.73$) but not between the beginning and the end of teacher education.

Relations of Process Characteristics to Prospective Preschool Teachers' Knowledge (H1)

As hypothesized, OTL in *general pedagogy* had a significant relation to GPK (H1a). If prospective preschool teachers reported more OTL in general pedagogy, they scored higher on the GPK assessment, and this occurred at a rate of 4.8 test points for each 1-point increase on the OTL scale (see Table 8). This corresponds to about half of a standard deviation, which is a medium effect size. As hypothesized, the relation between OTL in general pedagogy and GPK was also significantly stronger than the nonsignificant relation between OTL in mathematics and GPK. However, in contrast to our hypothesis the significant relation between OTL in mathematics pedagogy and GPK did not differ significantly from the effect of OTL in general pedagogy. OTL in general pedagogy were not significantly related to MPCK or MCK.

As hypothesized, OTL in mathematics pedagogy were strongly related to MPCK (H1b). The difference of 5.7 test points for a 1-point increase on the OTL scale represented about half a standard deviation and, thus, a medium effect size. As hypothesized, the relation was significantly stronger than the nonsignificant relation between OTL in mathematics and MPCK whereas it did not differ significantly from the relation between OTL in general pedagogy and MPCK. OTL in mathematics pedagogy were also significantly related to MCK and GPK, each time with small effect sizes of about 4 more test points for a 1-point increase on the OTL scale.

In contrast to H1c, OTL in mathematics did not have a significant relation to MCK. Only OTL in mathematics pedagogy were significantly related to this knowledge dimension.

Table 6
OTL In General Pedagogy (1), Mathematics Pedagogy (2), and Mathematics (3) Provided During Preschool Teacher Education

	Vocational school			Pedagogical college			*t* test for diff. between types of institutions			Cohen's *d*		
Variable	1	2	3	1	2	3	1	2	3	1	2	3
First year, *M* (*SD*)	2.91 (.34)	1.86 (.27)	1.81 (.40)	2.67 (.41)	1.79 (.68)	1.66 (.61)	−9.0*	−2.0^{ns}	−4.5*	.64	.14	.30
Last year, *M* (*SD*)	3.27 (.18)	2.30 (.42)	2.29 (.42)	3.22 (.20)	2.76 (.40)	2.27 (.32)	−3.3*	13.8*	−.82^{ns}	.26	1.12	.08
t test for diff. between program stages	25.4*	22.3*	21.5*	17.2*	17.9*	12.8*						
Cohen's *d*	1.33	1.28	1.17	1.80	1.80	1.31						

Note. OTL = opportunities to learn; *M* = mean; *SD* = standard deviation. Estimates were based on *t* tests of mean differences for independent samples.
* $p < .01$.

Table 7
Prospective Preschool Teachers' GPK (1), MPCK (2), and MCK (3) Scores

Variable	Vocational school			Pedagogical college			*t* test for diff. between types of institutions			Cohen's *d*		
	1	2	3	1	2	3	1	2	3	1	2	3
First year, *M* (*SD*)	48 (9.8)	47 (9.9)	48 (9.5)	51 (9.9)	52 (10.0)	55 (9.3)	4.6*	6.8*	10.4*	.31	.50	.74
Last year, *M* (*SD*)	50 (9.5)	50 (9.2)	48 (9.6)	57 (9.2)	58 (8.7)	55 (9.5)	9.7*	11.1*	8.8*	.74	.88	.73
Comparison												
t test for diff. between program stages	4.4*	5.7*	0.8^{ns}	7.2*	6.9*	0.2^{ns}						
Cohen's *d*	.21	.32	.00	.62	.63	.00						

Note. *M* = mean; *SD* = standard deviation of the person estimates from the 2PL IRT scaling. Estimates were based on *t* tests of mean differences for independent samples.
* $p < .01$.

The relations between OTL and the knowledge indicators played out the same way across all four subgroups, no matter whether the prospective preschool teachers were trained at vocational schools or in higher education or whether students were at the beginning or the end of their teacher education (H1d; see Table 9). Technically speaking, this means that freeing up the covariances between OTL and knowledge in the respective subgroups did not significantly improve the fit of the multiple-group model; Satorra-Bentler-scaled chi-square difference test (TRd): $\chi^2(8) = 3.7$, 9.0, or 4.9, respectively.

Relations of Structural Preschool Teacher Education Characteristics and Knowledge Outcomes (H2)

The hypothesis that the structural characteristics of preschool teacher education would predict prospective preschool teachers' knowledge was supported by the data (H2). The effect sizes were up to two thirds of a standard deviation (see Table 10). This applied to differential relations of types of institutions (H2a). Prospective preschool teachers from pedagogical colleges achieved significantly higher test scores than students from vocational schools, and this finding held with respect to GPK (+5.4), MPCK (+6.6), and MCK (+7.0).

The relations of the program stages were significant for GPK and MPCK. The knowledge in these two dimensions was higher at the end than at the beginning of preschool teacher education (H2b). The difference of about 3 test points corresponds to a small effect size. However, in contrast to our hypothesis, MCK was not higher at the end of preschool teacher education than it was at the beginning.

Table 8
Two-Level Models of Relations of OTL Provided During Preschool Teacher Education to Prospective Preschool Teachers' Knowledge (In Test Points)

Predictors	GPK	MPCK	MCK
OTL in general pedagogy	+4.8*	+1.8	−2.5
OTL in mathematics pedagogy	+3.7*	+5.7*	+4.3*
OTL in mathematics	−2.5	−1.8	−1.3

Note. GPK = general pedagogical knowledge; MPCK = mathematics pedagogical content knowledge; MCK = mathematics content knowledge; OTL = opportunities to learn. In all models, preschool teachers' gender, school marks in mathematics and German, language background, and mother's education were controlled for on the individual level.
* $p < .001$.

Mediation Models: Relations of Structure and Process to Knowledge (H3)

The relations of the structural preschool teacher education characteristics "type of institution" and "program stage" to GPK, MPCK, and MCK were hypothesized to be at least partly mediated through the respective domain-specific OTL (H3). The data supported this hypothesis with respect to program stage and all knowledge indicators as well as with respect to type of institution and MPCK. In contrast, the data did not support this hypothesis with respect to type of institution and GPK or MCK (see Figure 5). These unexpected findings suggest direct paths rather than indirect ones. The individual-level variables of gender, family background, and prior knowledge were controlled for in all three two-level models.

With respect to program stage, no significant direct relations to GPK, MPCK, or MCK existed any longer once the respective domain-specific OTL were included. The three knowledge indicators significantly depended on the extent to which OTL were provided in general pedagogy, mathematics pedagogy, or mathematics, which in turn significantly depended on whether the prospective preschool teachers were at the beginning or the end of their training (in favor of the latter). The additional indirect relation of this mediation of structure through process was significant for GPK and MPCK but not for MCK.

The picture differed with respect to the type of institution that provided preschool teacher education. In the case of MPCK was the relation of this structural characteristic partly mediated by OTL as a process characteristic. In addition to a significant direct relation of the type of institution to MPCK in favor of pedagogical colleges, a significant indirect relation existed. The OTL score in mathematics pedagogy was significantly related to MPCK, and this score in turn depended significantly on the type of institution (again in favor of pedagogical colleges).

By contrast, the type of institution did not matter significantly for how many OTL were offered in mathematics or in general pedagogy. So, in the cases of MCK and GPK, only significant direct relations of this structural characteristic to outcomes existed. The fit of all three mediation models was very good (GPK: CFI =

Table 9
Fit of Models for Testing Whether Relations of OTL To Outcomes Were Different in the Subgroups

	GPK			MPCK			MCK		
Model	LL	# par	BIC_{adj}	LL	# par	BIC_{adj}	LL	# par	BIC_{adj}
Equal	−7,257.8	13	14,571.2	−7,244.0	13	14,543.6	−7,221.2	13	14,498.0
Free	−7,254.1	21	14,597.9	−7,235.0	21	14,559.8	−7,216.3	21	14,522.5

Note. OTL = opportunities to learn; GPK = general pedagogical knowledge; MPCK = mathematics pedagogical content knowledge; MCK = mathematics content knowledge; LL = log likelihood; # Par = No. of parameters; BIC_{adj} = adjusted Bayesian information criterion. Equal = covariances constrained to be equal in the four subgroups, free = covariances estimated freely.

.98; RMSEA = .02; SRMR = .01 for the within model and .00 for the between model; MPCK: CFI = 1.00; RMSEA = .00, SRMR = .00 for the within and between models; MCK: CFI = .99; RMSEA = .01; SRMR = .01 for the within model and .00 for the between model).

Summary, Discussion, and Conclusions

Summary and Discussion

The relation of structural and process teacher education characteristics to outcomes was tested through multilevel modeling with 1,851 prospective preschool teachers nested in 86 classes from vocational schools and pedagogical colleges in Germany. Structural and process teacher education characteristics were modeled on the second level (classroom), whereas teacher background was controlled for on the first level (individual). The three knowledge indicators were modeled on both levels with data gathered through standardized and domain-specific testing, thus closing a much criticized gap in preschool research (Early et al., 2007; Whitebook et al., 2009).

The descriptive statistics revealed that OTL in mathematics and mathematics pedagogy were offered less often during preschool teacher education than OTL in general pedagogy confirming that the traditional concept of stronger emphasis on care than on cognitive development (Liegle, 2008) is still shaping preschool teacher education in Germany (for similar results in the US see Isenberg, 2000). This applied interestingly to programs at both types of institutions: vocational schools on the (post)secondary level and pedagogical colleges on the tertiary level—a result that may demonstrate that moving preschool teacher education up to the tertiary level alone may not be sufficient to change its nature. Prospective teachers from both types of institutions reported more OTL in general pedagogy already when they entered teacher education. This may go back to entrance requirements: the completion of vocational training in a care profession for vocational school students or a pedagogical internship for college students.

The data supported most but not all of our hypotheses. OTL in general pedagogy (H1a) and mathematics pedagogy (H1b) as well as types of institutions (H2a) and program stage (H2b) had significant relations to GPK or MPCK, respectively. Berliner's (1985) early call for a domain-specific perspective on teaching and learning contexts was thereby for the first time supported with respect to preschool teacher education in Germany.

The data also supported H3 that the effects of the distal structural preschool teacher education characteristics "type of institution" and "program stage" were partly mediated by OTL as proximal process characteristics. This applied in particular to program stage, and in the important case of MPCK, also to the type of institution. Process characteristics are obviously as crucial in the development of knowledge during preschool teacher education as structural characteristics. This result opens up for interesting con-

Table 10
Two-Level Models of Structural Preschool Teacher Education Effects on Prospective Preschool Teachers' Knowledge (in Test Points)

Predictors	GPK	MPCK	MCK
Type of institution	+5.4*	+6.6*	+7.0*
Program stage	+3.0*	+2.9*	−.2

Note. GPK = general pedagogical knowledge; MPCK = mathematics pedagogical content knowledge; MCK = mathematics content knowledge. Type of institution: 0 = vocational school, 1 = pedagogical college; program stage: 0 = beginning of teacher education, 1 = end of teacher education. In both models, preschool teachers' gender, school marks in mathematics and German, language background, and mother's education were controlled for on the individual level.
* $p < .01$.

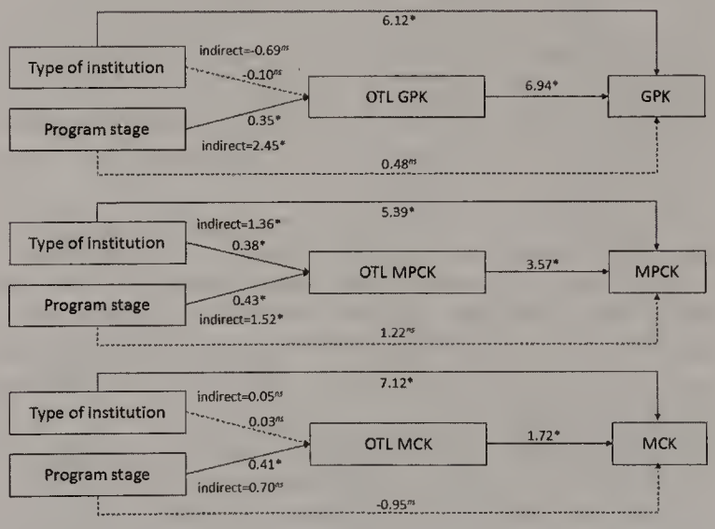


Figure 5. Mediation models of the relations between structural preschool teacher education characteristics, process characteristics, and outcomes (individual-level background variables are controlled for). OTL = opportunities to learn; GPK = general pedagogical knowledge; MPCK = mathematics pedagogical content knowledge; MCK = mathematics content knowledge. $p < .05$.

clusions beyond just moving programs from the secondary to the tertiary level (see below).

Outcomes of preschool teacher education in terms of GPK and MPCK already differed significantly at the beginning of the programs between vocational schools and pedagogical colleges in favor of the latter, which is probably an indicator of the stronger school credentials required by pedagogical colleges (graduation from high school instead of middle school). The differences between the beginning and the end of preschool teacher education were larger at pedagogical colleges than at vocational schools, a finding that can be interpreted as an indication of more OTL delivered during the longer programs at pedagogical colleges. Both results, the differences at the beginning and the differential development during teacher education, shed light on the much underresearched preschool teacher education below the tertiary level (Wallet, 2006). They point to severe disadvantages of this group of prospective teachers compared with college students.

OTL in mathematics and MCK behaved differently compared with MPCK and GPK. In contrast to our hypotheses H1c and H2b, neither a significant difference in MCK existed between the beginning and the end of teacher education, indicating a lack of progress during the programs, nor was there a significant relation between OTL in mathematics and MCK. Only OTL in mathematics pedagogy were significantly related to MCK. It seems as if OTL in mathematics pedagogy were better able to support MCK development although the correlational nature of the data asks for caution here.

Given that the content validity of the MCK test was confirmed in standardized expert reviews (Jenßen et al., 2015), the learning of MCK during preschool teacher education needs more research. It may be the case that the high degree of math anxiety found in a different sample of prospective preschool teachers played out negatively (Jenßen, Dunekacke, Eid et al., 2015) or that the more applied nature of mathematics in mathematics pedagogy OTL facilitated the acquisition of MCK for this group of teachers. If the latter result can be replicated, it has implications for preschool teacher education design and further research. In contrast to primary or secondary teachers where OTL in mathematics played a crucial role and MCK turned out to be a necessary prerequisite for MPCK (Blömeke et al., 2012; Tatto et al., 2012), the findings of the present study suggest that OTL in mathematics pedagogy may be more beneficial for acquiring MCK rather than OTL in mathematics.

This would be a unique result that distinguishes prospective preschool teachers from others. MPCK may build more appropriately on preschool teachers' prior knowledge because students come into the programs with more experience in general pedagogy. Learning typically happens through connecting new information to prior knowledge (Carroll, 1963). OTL in mathematics pedagogy may elicit teachers' prior knowledge and help them to make connections to new knowledge, thus serving as a bridge and therefore being an effective instructional approach.

Before conclusions are drawn from these results and interpretations, some methodological limitations of the study need to be pointed out. Institutions provide a set of intertwined organizational and pedagogical characteristics (Tinto, 1998) so that other characteristics than those examined here could be causal (e.g., climate or composition effects). It is not possible to disentangle such effects in a correlational study. Furthermore, the fuzzy notion of "OTL" which has been defined in different ways in educational research leads to a lot of noise in the state of research. The results would therefore be substantially strengthened if they were replicated with other samples. Other researchers in the field are requested to apply the OTL measures and knowledge tests in further samples, in particular in other countries, or to develop new measures based on the same construct definitions so that cross-validations can take place. This is necessary to avoid acting too quickly on the basis of one study that was conducted in only one national context.

Conclusions

Given the small amount of OTL offered in mathematics pedagogy as indicated by the descriptive statistics, the findings presented in this paper have to be of concern with respect to fostering children's mathematics literacy (Duncan et al., 2007; Reynolds, 1995). Even though preschool standards require preschool teachers to achieve ambitious objectives (Jenßen et al., 2015), prospective teachers do not seem to receive sufficient OTL. Because of the interdisciplinary nature of MPCK as an amalgam of MCK and GPK (Shulman, 1986), this dimension of knowledge seems to be crucial for the development of teacher knowledge in general. The significant relation between OTL in mathematics pedagogy and outcomes indicates that it may be worthwhile to increase the number of OTL in this domain, particularly at vocational schools (Janssen, 2010).

Note that the conclusion that more OTL should be provided differs from a request to train all preschool teachers at the higher education level. The finding that processes mediate structural characteristics may open up new ways of thinking in this context. If it is OTL that count, these can also be provided at vocational schools. However, differential intake characteristics have still to be taken into account. Furthermore, the colleges have the advantage that their preschool teacher education program lasts for 4 years instead of only 2, thus providing more teaching time to prepare preschool teachers for all the tasks they have to cover. The question of how to provide the higher costs coming with longer preschool teacher education has to be taken into account in this context, too.

Besides such policy-related conclusions, a broad range of research needs can be derived from the present study. Future studies should address mediations through cross-domain OTL in particular with respect to mathematics pedagogy and general pedagogy. Furthermore, it needs to be examined how preschool teachers knowledge base is transformed into job performance because ultimately preschool teacher education is designed to support high quality teaching in preschool and child development. Not many studies have taken on examining such long-term effects of teacher education because this needs sophisticated designs and standardized observation protocols or video-based measures of teaching skills. Although difficult to implement, such research would provide urgently needed information on how to structure preschool teacher education so that prospective teachers acquire an appropriate knowledge and skill base to succeed in their job.

References

Abell Foundation. (2001). *Teacher certification reconsidered: Stumbling for quality*. Baltimore, MD: Author.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA.

Anders, Y. (2012). *Modelle professioneller Kompetenzen für frühpädagogische Fachkräfte: Aktueller Stand und ihr Bezug zur Professionalisierung* [Models of professional competencies of early childhood education teachers: State of research and its impact on professionalization]. München, Germany: Knoblingesign.

Anderson, L., & Krathwohl, D. A. (2001). *Taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42,* I7–I16. http://dx.doi.org/10.1097/01.mlr.0000103528.48582.7c

Bassok, D. (2012). *Raising teacher education levels in Head Start: Are there program-level tradeoffs?* (CEPWC Working Paper; 3). Charlottesville, VA: Center on Education Policy and Workforce Competitiveness.

Berliner, D. C. (1985). Effective classroom teaching: The necessary but not sufficient condition for developing exemplary schools. In G. R. Austin & H. Garber (Eds.), *Research on exemplary schools* (pp. 127–154). Orlando, FL: Academic Press Ins. http://dx.doi.org/10.1016/B978-0-12-068590-5.50013-9

Blömeke, S., Jenßen, L., Dunekacke, S., Suhl, U., Grassmann, M., & Wedekind, H. (2015). Professionelle Kompetenz von Erzieherinnen messen: Entwicklung und Validierung standardisierter Leistungstests für frühpädagogische Fachkräfte [Assessment of professional competence of preschool teachers: Development and validation of standardized achievement tests for early childhood personell]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 29,* 177–191.

Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.). (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer: Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare—Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung* [Professional competencies of future teachers: Knowledge, beliefs and opportunities to learn of German mathematics teachers–First results on the effectiveness of teacher education]. Münster, Germany: Waxmann.

Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education, 62,* 154–171. http://dx.doi.org/10.1177/0022487110386798

Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: What matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education, 28,* 44–55. http://dx.doi.org/10.1016/j.tate.2011.08.006

Bogard, K., Traylor, F., & Takanishi, R. (2008). Teacher education and PK outcomes: Are we asking the right questions? *Early Childhood Research Quarterly, 23,* 1–6. http://dx.doi.org/10.1016/j.ecresq.2007.08.002

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling, 19,* 372–398. http://dx.doi.org/10.1080/10705511.2012.687671

Burchinal, M. R., Cryer, D., Clifford, R. M., & Howes, C. (2002). Caregiver training and classroom quality in child care centers. *Applied Developmental Science, 6,* 2–11. http://dx.doi.org/10.1207/S1532480XADS0601_01

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16,* 572–578.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64,* 722–733.

Clements, D. H., Sarama, J., & DiBiase, A-M. (Eds.). (2004). *Engaging young children in mathematics: Standards for early childhood mathematics*. Mahwah, NJ: Erlbaum.

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology, 100,* 1261–1293. http://dx.doi.org/10.1086/230638

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. http://dx.doi.org/10.1177/001316446002000104

Connor, C. M., Morrison, F. J., & Slominski, L. (2006). Preschool instruction and children's emergent literacy growth. *Journal of Educational Psychology, 98,* 665–689. http://dx.doi.org/10.1037/0022-0663.98.4.665

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*. Retrieved from http://epaa.asu.edu/ojs/article/view/392/515.

Dawis, R. V. (2000). Scale construction and psychometric considerations. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 65–94). Orlando, FL: Academic Press. http://dx.doi.org/10.1016/B978-012691360-6/50004-5

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428

Dunekacke, S., Jenßen, L., & Blömeke, S. (2015a). Effects of mathematics content knowledge on pre-school teachers' performance: A video-based assessment of perception and planning abilities in informal learning situations. *International Journal of Science and Mathematics Education, 13,* 267–286. http://dx.doi.org/10.1007/s10763-014-9596-z

Dunekacke, S., Jenßen, L., & Blömeke, S. (2015b). Mathematikdidaktische Kompetenz von Erzieherinnen und Erziehern: Validierung des KomMa-Leistungstests durch die videogestützte Erhebung von Performanz [Competencies in mathematics pedagogy of preschool teachers: Validation of the KomMa Achievement Test through video-based assessments of performance]. *Zeitschrift fur Padagogik, 61,* 80–99.

Early, D. M., Bryant, D. M., Pianta, R. C., Clifford, R. M., Burchinal, M. R., Ritchie, S., . . . Barbarin, O. (2006). Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten? *Early Childhood Research Quarterly, 21,* 174–195. http://dx.doi.org/10.1016/j.ecresq.2006.04.004

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., . . . Zill, N. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78,* 558–580. http://dx.doi.org/10.1111/j.1467-8624.2007.01014.x

Early, D. M., & Winton, P. J. (2001). Preparing the workforce: Early childhood teacher preparation at 2- and 4-year institutions of higher education. *Early Childhood Research Quarterly, 16,* 285–306. http://dx.doi.org/10.1016/S0885-2006(01)00106-5

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50,* 328–344.

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., . . . Scott-Little, C. (2012). A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal, 49,* 88–123. http://dx.doi.org/10.3102/0002831211434596

Howes, C., Whitebook, M., & Phillips, D. (1992). Teacher characteristics and effective teaching in child care: Findings from the National Child Care Staffing Study. *Child & Youth Care Forum, 21,* 399–414. http://dx.doi.org/10.1007/BF00757371

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Isenberg, J. P. (2000). The state of the art in early childhood professional preparation. In *New teachers for a new century: The future of early childhood professional preparation* (pp. 17–58). Washington, DC: National Institute on Early Childhood Development and Education, U.S. Department of Education.

Janssen, R. (2010). *Die Ausbildung Frühpädagogischer Fachkräfte an Berufsfachschulen und Fachschulen: Eine Aualyse im Ländervergleich* [Teacher education of preschool teachers at different types of vocational schools at the secondary level: An analysis across states]. Munich, Germany: DJI.

Jenßen, L., Dunekacke, S., Baack, W., Tengler, M., Koinzer, T., Schmude, C., . . . Blömeke, S. (2015). KomMa: Kompetenzmodellierung und Kompetenzmessung bei frühpädagogischen Fachkräften im Bereich Mathematik [KomMa: Modelling and assessing competencies of preschool teachers in mathematics]. In B. Koch-Priewe, A. Köker, J. Seifried, & E. Wuttke (Eds.), *Kompetenzerwerb an Hochschulen: Modellieruug und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte* [Competence acquisition at universities: Modelling and measurement. To the professionalization of beginning teachers and teachers as well asearly-educational specialists] (pp. 59–79). Bad Heilbrunn, Germany: Klinkhardt.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis [Quality assurance in research on competencies: Suggestions for the validation of test development and publication practices]. *Zeitschrift fur Padagogik, 61,* 11–31.

Jenßen, L., Dunekacke, S., Eid, M., & Blömeke, S. (2015). The relationship of mathematical competence and mathematics anxiety: An application of latent state-trait theory. *Zeitschrift fur Psychologie mit Zeitschrift fur Angewandte Psychologie, 223,* 31–38. http://dx.doi.org/10.1027/2151-2604/a000197

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426. http://dx.doi.org/10.1007/BF02291366

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., & Baumert, J. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of Teacher Education, 64,* 90–106. http://dx.doi.org/10.1177/0022487112460398

Klusmann, U., Kunter, M., Voss, T., & Baumert, J. (2012). Berufliche Beanspruchung angehender Lehrkräfte: Die Effekte von Persönlichkeit, pädagogischer Vorerfahrung und professioneller Kompetenz [Emotional exhaustion and job satisfaction of beginning teachers: The role of personality, educational experience and professional competence]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 26,* 275–290. http://dx.doi.org/10.1024/1010-0652/a000078

König, J., Blömeke, S., Paine, L., Schmidt, W. H., & Hsieh, F.-J. (2011). General pedagogical knowledge of future middle school teachers: On the complex ecology of teacher education in the United States, Germany, and Taiwan. *Journal of Teacher Education, 62,* 188–201. http://dx.doi.org/10.1177/0022487110388664

Landry, S. H., Anthony, J. L., Swank, P. R., & Monseque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology, 101,* 448–465. http://dx.doi.org/10.1037/a0013842

Liegle, L. (2008). Erziehung als Aufforderung zur Bildung: Aufgaben der Fachkräfte in Tageseinrichtungen für Kinder in der Perspektive der frühpädagogischen Didaktik [Education as a request of forming oneself: Demands preschool teachers are confronted with from a didactical point of view]. In W. Thole, H.-G. Roβbach, M. Fölling-Albers, & R. Tippelt (Eds.), *Bildung und Kindheit. Pädagogik der Frühen Kindheit in Wissenschaft und Lehre* [Education and childhood. Pedagogy of the early childhood in science and teachings] (pp. 85–113). Opladen, Germany: Budrich.

Lobman, C., Ryan, Sh. & McLaughlin, J. (2005). Reconstructing teacher education to prepare qualified preschool teachers: Lessons from New Jersey. *Early Childhood Research & Practice, 7,* 30–35. Retrieved from http://ecrp.uiuc.edu/v7n2/lobman.html

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau, 58,* 103–117. http://dx.doi.org/10.1026/0033-3042.58.2.103

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education, 30,* 520–540. http://dx.doi.org/10.2307/749772

Muthén, L. K., & Muthén, B. O. (2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

NAEYC. (2009). *NAEYC Standards for early childhood professional preparation position statement approved by the NAEYC Governing Board July 2009.* Washington, DC: Author.

National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity.* Washington, DC: National Academies Press.

OECD. (2004). *Starting strong: Curricula and pedagogies in early childhood education and care.* Paris, France: Author.

Pianta, R. C., DeCoster, J., Cabell, S., Burchinal, M., Hamre, B. K., Downer, J., . . . Howes, C. (2014). Dose–response relations between preschool teachers' exposure to components of professional development and increases in quality of their interactions with children. *Early Childhood Research Quarterly, 29,* 499–508. http://dx.doi.org/10.1016/j.ecresq.2014.06.001

Piasta, S. B., Logan, J. A. R., Pelatti, C. Y., Capps, J. L., & Petrill, S. A. (2015). Professional development for early childhood educators: Efforts to improve math and science learning opportunities in early childhood classrooms. *Journal of Educational Psychology, 107,* 407–422. http://dx.doi.org/10.1037/a0037621

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74,* 423–452.

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling, 17,* 122–132. http://dx.doi.org/10.1080/10705511003659417

Reynolds, A. (1995). One year of preschool intervention or two: Does it matter? *Early Childhood Research Quarterly, 10,* 1–31. http://dx.doi.org/10.1016/0885-2006(95)90024-1

Rowan, B., Correnti, R., & Miller, R. J. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of Elementary Schools* (CPRE Research Report Series RR-051). Philadelphia, PA: University of Pennsylvania.

Rubin, D. (1976). Inference and missing data. *Biometrika, 63,* 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

Ryan, S. H., Ackerman, D. J., & Song, H. (2004). *Getting qualified and becoming knowledgeable: Preschool teachers' perspectives on their professional preparation.* Rutgers, NJ: State University of New Jersey.

Scheerens, J., & Blömeke, S. (2016). Integrating teacher education effectiveness research into educational effectiveness models. *Educational Research Review, 18,* 70–87. http://dx.doi.org/10.1016/j.edurev.2016.03.002

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15,* 4–14. http://dx.doi.org/10.3102/0013189X015002004

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Statistisches Bundesamt. (2010). *Frauen und Männer in verschiedenen Lebensphasen* [Men and women at different life-stages]. Wiesbaden, Germany: Statistisches Bundesamt.

Statistisches Bundesamt. (2014). *Statistiken der Kinder- und Jugendhilfe: Kinder und tätige Personen in Tageseinrichtungen und in öffentlich geförderter Kindertagespflege am 01.03.2014* [Statistics of early childhood education and care: Children and employees at day-care centres and publicly funded day care on March 1, 2014]. Wiesbaden, Germany: Author.

Tatto, M. T., Schwille, J., Senk, S. L., Ingvarson, L., Rowley, G., Peck, R., . . . Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and development Study in Mathematics (TEDS-M).* Amsterdam, the Netherlands: IEA.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research.* London, UK: Falmer Press.

Tinto, V. (1998). Colleges as communities: Taking research on student persistence seriously. *The Review of Higher Education, 21,* 167–177.

Tout, K., Zaslow, M., & Berry, D. (2005). Quality and qualifications: Links between professional development and quality in early care and education settings. In M. Zaslow & I. Martinez-Beck (Eds.), *Critical issues in early childhood professional development* (pp. 77–110). Baltimore, MD: Paul H. Brooks.

Travers, K. J., & Westbury, I. (1989). *The IEA study of mathematics 1: Analysis of mathematics curricula.* Oxford, UK: Pergamon Press.

Voss, Th., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology, 103,* 952–969. http://dx.doi.org/10.1037/a0025125

Wallet, P. (2006). *Pre-primary teachers: A global analysis of several key education indicators.* Paper commissioned for the EFA Global Monitoring Report 2007 "Strong foundations: Early childhood care and education" (2007/ED/EFA/MRT/PI/32). Paris, France: UNESCO.

Weinert, F. E. (2001). Concept of competence: A conceptual classification. In D. S. Rychen & L. Hersh Salganik (Eds.), *Defining and selecting key competencies.* Göttingen, Germany: Hogrefe.

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 84–136). San Francisco, CA: Jossey-Bass. http://dx.doi.org/10.2307/270754

Whitebook, M., Gomby, D., Bellm, D., Sakai, L., & Kipris, F. (2009). *Preparing teachers of young children: The current state of knowledge, and a blueprint for the future.* Berkeley, CA: Center for the Study of Child Care Employment.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75,* 579–642. http://dx.doi.org/10.2307/2340126

(Appendix follows)

Appendix

Dimensions, Subdimensions, and Descriptors of Preschool Teacher Knowledge (in Parentheses: No. of Test Items)

General pedagogical knowledge (18)	Mathematics pedagogical content knowledge (28)	Mathematics content knowledge (24)
Educational foundations (5) – Knowledge of fundamental educational terms (1) – Selection of educational objectives for children aged 3–6 (1) – Application of educational approaches (2) – Formal and informal opportunities to learn (1)	Diagnosing children's mathematical development (17) – Developmental psychology of children's mathematical competencies (2) – Diagnosing developmental states in the field of number, sets, and operations based on children's statements (5) – Diagnosing developmental states in the field of shape, space, and change based on children's statements (2) – Evaluation of standardized and unstandardized diagnostic approaches (2) – Identification of everyday-life situations with relations to numbers, sets, and operations (3) – Identification of everyday-life situations with relations to shape, space, and change (1) – Identification of everyday-life situations with relations to quantity, measurement, and relations (2)	Numbers, sets, and operations (6) – Knowledge of number range (1) – Application of basic operations (2) – Application of number principles (2) – Understanding sets (1) Shape, space, and change (6) – Application of formulas (2) – Recognizing geometrical shapes (2) – Demonstrating space orientation (1) – Constructing geometrical shapes (1)
Psychological foundations (6) – Knowledge of motivation and attribution theories (2) – Diagnosing general learning and developmental processes of 3-to-6-year-old children (2) – Development of strategies to change child behavior (2)		
Instructional foundations (7) – Application of communication and collaboration approaches (2) – Application of approaches to foster learning and development in heterogeneous groups of children between the ages of 3 and 6 (3) – Application of inclusive principles (2)	Designing informal learning environments that foster mathematical learning (11) – Application of approaches that support mathematical learning (incl. specifics for children at risk) (3) – Initiate play-based experiences with numbers, sets, and operations (2) – Initiate play-based experiences with shape, space, and change (4) – Initiate play-based experiences with data, combinatorics, and chance (1) – Initiate play-based experiences with quantity, measurement, and relations (1)	Data, combinatorics, and chance (6) – Generating tables and lists of frequencies (2) – Estimating the number of possibilities (1) – Estimating chance (3) Quantity, measurement, and relations (6) – Relating speed to time (2) – Transforming verbal into mathematical statements and vice versa (2) – Pattern recognition (2)

Received November 22, 2015
Revision received June 29, 2016
Accepted June 29, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 3, 355–373

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000145

Supporting Students in Making Sense of Connections and in Becoming Perceptually Fluent in Making Connections Among Multiple Graphical Representations

Martina A. Rau
University of Wisconsin, Madison

Vincent Aleven
Carnegie Mellon University

Nikol Rummel
Ruhr-Universität Bochum

Prior research shows that multiple representations can enhance learning, provided that students make connections among them. We hypothesized that support for connection making is most effective in enhancing learning of domain knowledge if it helps students both in making sense of these connections and in becoming perceptually fluent in making connections. We tested this hypothesis in an experiment with 428 4th- and 5th-grade students who worked with different versions of an intelligent tutoring system for fractions learning. Results did not show main effects for sense-making or fluency-building support but an interaction effect, such that a combination of sense-making and fluency-building support is most effective in enhancing fractions knowledge. Causal path analysis of log data from the system shows that sense-making support enhances students' benefit from fluency-building support, but fluency-building support does not enhance their benefit from sense-making support. Our results suggest that both understanding of connections and perceptual fluency in connection making are critical aspects of learning of domain knowledge with multiple graphical representations. Findings from the causal path analysis lead to the testable prediction that instruction should provide sense-making support and fluency-building support for connection making.

Keywords: multiple representations, connection making, intelligent tutoring systems, classroom evaluation, causal path analysis

Supplemental materials: http://dx.doi.org/10.1037/edu0000145.supp

Instructional materials typically use a variety of representations. For instance, students learning about fractions usually encounter the representations shown in Figure 1: circles, rectangles, and number lines. There is considerable evidence for benefits of multiple representations on students' learning (Ainsworth, 2006; de Jong et al., 1998; Eilam & Poyas, 2008). Multiple representations can enhance learning because they emphasize complementary conceptual aspects of the content (Larkin & Simon, 1987; Schnotz, 2005; Schnotz & Bannert, 2003). For example, the circle in Figure 1 depicts fractions as part of a whole circle, whereas the number line depicts fractions as a measure of length.

However, students' benefit from multiple representations depends on their ability to make connections among them (Ainsworth, 2006; Cook, Wiebe, & Carter, 2008; Taber, 2001). For example, learning of fractions requires an integration of the different concepts afforded by the representations in Figure 1 (National Mathematics Advisory Panel, 2008; Siegler et al., 2010). Therefore, students need to make connections among these representations. However, connection making is a difficult task (de Jong et al., 1998; Van Someren, Boshuizen, & de Jong, 1998) that students often fail to attempt spontaneously (Ainsworth, Bibby, & Wood, 2002; Rau, Aleven, Rummel, & Pardos, 2014). At least two types of connection-making competencies play a role in students' learning. First, they need *understanding of connections*: the ability to map corresponding visual features of the graphical representations (GRs) to one another (e.g., Ainsworth, 2006; Schnotz & Bannert, 2003; Seufert, 2003). For example, when working with the GRs in Figure 1, students may map the colored section in the circle to the number of sections between 0 and the dot in the number line, based on the rationale that both show the numerator of the fraction. Second, connection making involves the acquisition of *perceptual fluency:* learning to recognize visual patterns in

This article was published Online First August 8, 2016.

Martina A. Rau, Department of Educational Psychology, University of Wisconsin, Madison; Vincent Aleven, Human-Computer Interaction Institute, Carnegie Mellon University; Nikol Rummel, Institute of Educational Research, Ruhr-Universität Bochum.

This work was supported by the National Science Foundation, REESE-21851-1-1121307, by the IES R305A120734, and by the PSLC, funded by NSF award number SBE-0354420. We thank Richard Scheines, Ken Koedinger, Mitchell Nathan, Kathy Cramer, Jay Raspat, Michael Ringenberg, Brian Junker, Howard Seltman, Cassandra Studer, the students, teachers, and principals, the Tetrat, CTAT, and Datashop teams, especially Mike Komisin, Alida Skogsholm, and Joseph Ramsey.

Correspondence concerning this article should be addressed to Martina A. Rau, Department of Educational Psychology, University of Wisconsin, 1025 West Johnson Street, Madison, WI 53706. E-mail: marau@wisc.edu

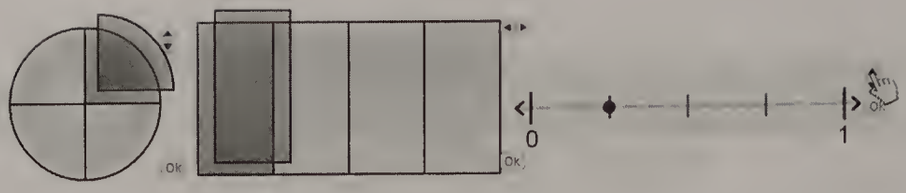


Figure 1. Graphical representations of fractions: circle, rectangle, and number line. See the online article for the color version of this figure.

GRs that correspond to domain-relevant concepts. For example, the student may learn to recognize that the GRs in Figure 1 show the same proportion of some unit.

Although prior research has yielded a number of effective interventions to support both types of connection-making competencies, this research has so far not investigated possible interactions among them. Our work addresses this gap by investigating whether combining support tailored to each type of connection-making competency enhances students' learning of fractions knowledge. We chose fractions as a domain for our research because—similar to many other STEM domains—instructional materials typically use multiple graphical representations (MGRs) that emphasize different concepts. Therefore, our research has the potential to generalize to other STEM domains. We conducted our research as part of regular classroom instruction in the context intelligent tutoring systems (Koedinger & Corbett, 2006), which are used in many classrooms across the United States and hence represent a realistic educational scenario. A further advantage of intelligent tutoring systems is that they allow for the use of interactive, virtual GRs while providing tutoring, which aligns with mathematics education research demonstrating advantages of virtual over physical GRs for fractions instruction (Moyer, Bolyard, & Spikell, 2002; Reimer & Moyer, 2005). For our experiment, we used the Fractions Tutor (Rau, Aleven, Rummel, & Rohrbach, 2013), which provides multiple virtual GRs and has been shown to yield significant learning about fractions knowledge among elementary-school students.

Motivation

Multiple Graphical Representations of Fractions

The mathematics education literature suggests that GRs fundamentally shape how students conceptualize fractions (Charalambous & Pitta-Pantazi, 2007; Cramer, Wyberg, & Leavitt, 2008). Fractions are a notoriously complex topic (Charalambous & Pitta-Pantazi, 2007). Indeed, Behr, Lesh, Post, and Silver (1993) suggest at least six conceptual ways to interpret fractions: (a) parts of a whole, (b) decimals, (c) ratios, (d) quotient, (e) operators, and (f) measurements. GRs differ in their capacity to help students understand these concepts. For instance, area models (i.e., circles and rectangles) can illustrate part-whole concepts (e.g., one of four sections is shaded), ratio concepts (one section is shaded, three are unshaded), and quotient concepts (one whole divided by four; Cramer et al., 2008). While circles are a type of area model in which the whole is inherent in the shape (i.e., a full circle; Cramer et al., 2008), rectangles do not have a standard shape but can be divided horizontally and vertically, which is helpful for illustrating quotient and operator interpretations. By contrast, linear models (e.g., number lines) are well suited to illustrate measurement and decimal concepts (Siegler et al., 2010).

Fractions instruction typically uses multiple-graphical-representations (MGRs; Charalambous & Pitta-Pantazi, 2007; Kieren, 1993; Lamon, 1999; Martinie & Bay-Williams, 2003; Moss & Case, 1999; Thompson & Saldanha, 2003). Common curricula tend to start fractions instruction with area models (e.g., circles and rectangles) to introduce part-whole concepts of fractions and then work toward including other concepts, for instance by using number lines to illustrate measurement concepts (Behr et al., 1993; Kieren, 1993; Ohlsson, 1988). Failure to make connections among these different GRs may lead students to overly rely on one conceptual interpretation (Behr et al., 1993; Kieren, 1993; Ohlsson, 1988). This can cause misconceptions such as the "whole number bias": the bias to treat fractions as composites of whole numbers (i.e., numerator and denominator), rather than as overall fraction values (Ni & Zhou, 2005). Indeed, Siegler and colleagues criticize early reliance on area models in fractions instruction for overemphasizing part-whole concepts (Siegler et al., 2011, 2013). Instead, they recommend increased use of number line representations to emphasize measurement concepts. In line with this recommendation, educational practice guides emphasize advantages of number lines over other GRs (National Mathematics Advisory Panel, 2008; Siegler et al., 2010).

Given recent research on the potential privilege of number line representations over area models (Siegler et al., 2011, 2013), one may even argue that *unless* students make connections among GRs, they may learn better with number lines alone. Indeed, in our own prior research, we found that students benefited from MGRs only if they received instructional support to relate each GR to key fractions concepts (Rau, Aleven, & Rummel, 2015). Without this support, students who worked with number lines alone showed higher learning gains than students who worked with MGRs.

In summary, students' benefit from MGRs depends on their ability to make connections among them. However, it remains an open question how best to support students in making such connections. We investigate this question in our current experiment. Because Siegler's suggestion that number lines alone may be more effective than MGRs is mainly rooted in concerns about failure to connect measurement concepts to part-whole concepts, our experiment focuses on connection making between the GRs typically used to emphasize these concepts: number lines and area representations (circle and rectangle).

Theoretical Framework

To address the question of how best to support students in making connections among MGRs, we draw on a recent theoretical framework that seeks to bridge cognitive science and educational

research to educational practice: Koedinger and colleagues' (2012) Knowledge-Learning-Instruction framework (KLI; also see Koedinger et al., 2013). KLI offers the *alignment hypothesis:* instructional interventions are most effective if they enhance learning processes that match the complexity of the to-be-learned competency. Hence, we use KLI to consider (a) the complexity of connection-making competencies that are important for domain expertise, (b) through which learning processes students acquire these competencies, and (c) which instructional interventions may match their complexity. As illustrated in Figure 2, these theoretical considerations lead to the hypothesis that combining support tailored to each type of connection-making competency enhances students' learning of fractions knowledge.

Connection-making competencies in domain expertise. The literature on expertise provides insights into how connection making among MGRs relates to domain expertise. Our review of this research suggests that two connection-making competencies play an important role in expertise (see Rau, 2016, for an overview): *understanding* of connections (Ainsworth, 2006; Dreyfus & Dreyfus, 1986; Patel & Dexter, 2014; Richman et al., 1996) and *perceptual fluency* in connection making (Dreyfus & Dreyfus, 1986; Gibson, 1969, 2000; Pape & Tchoshanov, 2001; Richman et al., 1996). To analyze the complexity of these competencies, we draw on KLI's definition of a *knowledge component* as an "acquired unit of cognitive function . . . that can be inferred from performance on a set of related tasks" (Koedinger et al., 2012, p. 764).

Understanding of connections among GRs means that a student can map visual features of one GR to those of a different GR because they show the same concept (Ainsworth, 2006; Charalambous & Pitta-Pantazi, 2007; Cramer, 2001; Kozma & Russell, 2005; Patel & Dexter, 2014). For example, consider a student who sees the GRs shown in Figure 1. The student may map the shaded section in the circle to the section between zero and the dot in the number line because both visual features depict the numerator, and he or she may relate the number of total sections in the circle to the sections between 0 and 1 in the number line because both features show the denominator. By reasoning about these connections, the student may understand the abstract principle that both GRs express fractions as portions of a unit, measured by partitioning the unit into equal sections. Under KLI, such reasoning involves learning of *complex* knowledge components because it requires that students learn a principle that applies in multiple situations (e.g., a proportion can be shown in multiple ways: circles, rectangles, number lines, etc.).

Perceptual fluency in making connections is the ability to quickly and effortlessly see holistic, corresponding visual patterns across different GRs. For example, a student should see "at a glance" that the circle and the number line show the same proportion of a unit. Perceptual fluency in connection making is related to domain expertise because it frees "cognitive head room" that allows students to reason about domain-relevant concepts (Gibson, 2000; Kellman & Massey, 2013; Richman et al., 1996). Under KLI, perceptual fluency involves learning of *simple* knowledge components because there is a one-to-one mapping between the GRs (e.g., circle and number line) and the visual pattern (e.g., proportion of unit covered).

Connection-making processes that lead to connection-making competencies. KLI moves beyond the analysis of knowledge components by relating them to the learning processes through which students acquire them. Students learn complex knowledge components via *sense-making processes*. These processes are verbally mediated because they involve explanations of principles of how GRs depict conceptually relevant information (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Gentner, 1983; Koedinger et al., 2012). They are explicit in that students have to willfully engage in them (Chi, de Leeuw, Chiu, & Lavancher, 1994; diSessa & Sherin, 2000). The literature on learning with

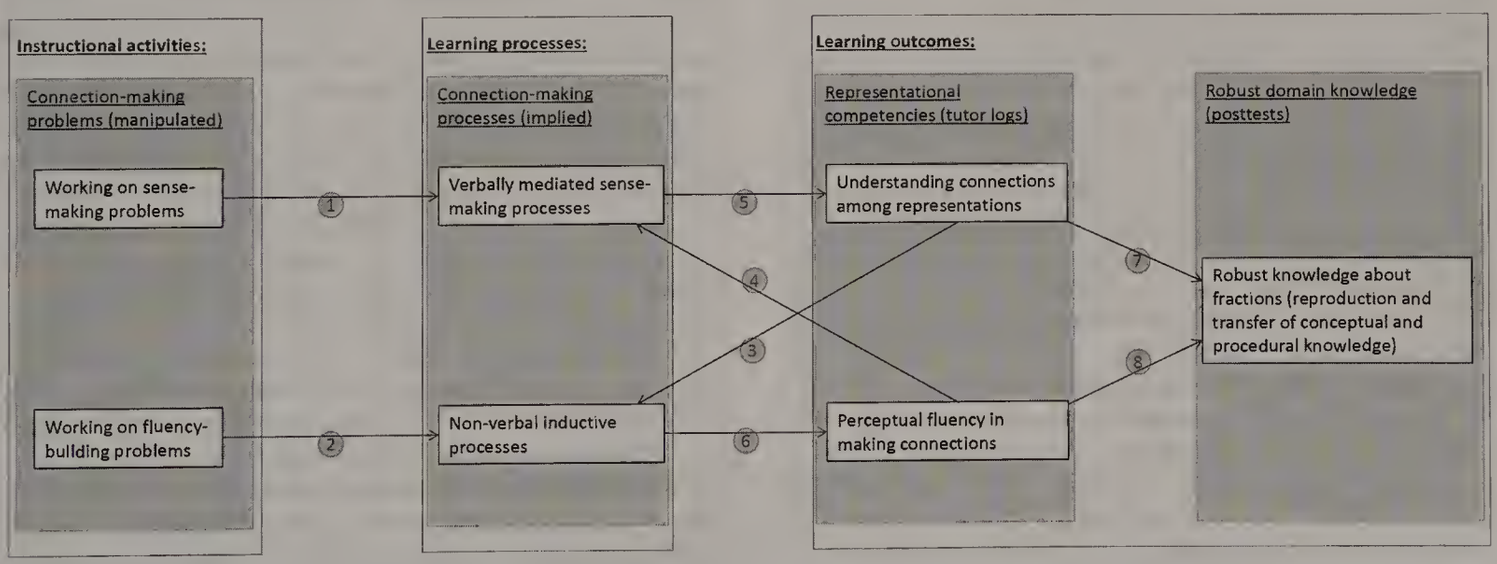


Figure 2. Theory of change of how working on connection-making problems (sense-making problems, fluency-building problems) foster learning processes (verbally mediated sense-making processes, nonverbal inductive and refinement processes) and representational competences (understanding of connections and perceptual fluency in making connections) that enhance students' learning of robust domain knowledge (robust fractions knowledge). For each mechanism, the figure indicates which section in the article describes prior research regarding this particular mechanism. See the online article for the color version of this figure.

representations often refers to sense-making processes as structure mapping processes (Gentner & Markman, 1997) because students map features of the representations to abstract concepts. Seufert (2003) suggests that structure mapping is one major process through which students integrate information from multiple representations into a coherent understanding of domain knowledge. diSessa's (2004) framework of metarepresentational competence and research on representational flexibility (Acevedo Nistal, Van Dooren, & Verschaffel, 2013, 2015) suggest that sense-making processes are also involved in selecting appropriate GRs to solve domain-relevant problems.

By contrast, students learn simple knowledge components via *nonverbal inductive learning processes* (Koedinger et al., 2012; Richman et al., 1996) that they engage in when learning to categorize instances accurately and efficiently (Koedinger et al., 2012). These processes are often nonverbal because they do not require explicit reasoning (Kellman & Garrigan, 2009; Kellman & Massey, 2013). They are implicit because they typically happen unintentionally and unconsciously (Shanks, 2005) through experience with many instances (Gibson, 1969, 2000; Kellman & Massey, 2013; Richman et al., 1996). The literature also refers to inductive learning processes as perceptual learning and pattern recognition (Gibson, 1969; Goldstone & Barsalou, 1998; Kellman & Massey, 2013; Richman et al., 1996).

Instructional interventions to support connection-making processes. According to KLI's *alignment hypothesis*, instructional interventions that enhance sense-making processes are most effective for complex knowledge components, whereas interventions that enhance inductive processes are most effective for simple knowledge components.

Supporting verbally mediated sense-making processes in connection making. KLI identifies principles that can guide the design of instructional activities that support sense-making processes (Koedinger et al., 2013). Here, we discuss two instructional activities that apply to the case of connection making: *explicitly comparing* multiple instances and providing *self-explanation prompts*. Prior research has demonstrated how best to implement these principles into support for connection making. First, sense-making support is particularly effective if it prompts students to self-explain mappings between representations (Ainsworth & van Labeke, 2002; Bodemer & Faust, 2006; Seufert, 2003; van der Meij & de Jong, 2011). Such prompts may be critical because students typically struggle in making sense of connections (Ainsworth et al., 2002), especially if they have low prior knowledge (Stern, Aprea, & Ebner, 2003). For example, Berthold and Renkl (2009) show that self-explanation prompts increase students' benefit from multiple representations. In their experiment, self-explanation prompts were implemented in the form of "why"-questions, to elicit self-explanations of principled knowledge. Self-explanation prompts are more effective if they ask students to self-explain specific connections than if they are open-ended (Berthold, Eysink, & Renkl, 2008; van der Meij & de Jong, 2011).

Second, sense-making support typically asks students to use these mappings to compare how representations show analogous information or different, complementary information about the concepts they depict (Bodemer & Faust, 2006; Seufert, 2003; Seufert & Brünken, 2006; van der Meij & de Jong, 2006; Van Labeke & Ainsworth, 2002; Vreman-de Olde & De Jong, 2006). Although most implementations of sense-making support encourage students to compare representations, our review of prior research showed that there are two different, commonly used implementations. One common implementation of sense-making support in computer-based learning environments uses *linked representations*, where the student's manipulations of one GR are automatically reflected in the other GR (e.g., Ainsworth & van Labeke, 2002; van der Meij & de Jong, 2006, 2011). Linked GRs allow students to explore intermediate steps, mistakes, and the final result in two or more GRs. This implementation aligns with KLI's cognitive dissonance principle, which states that presenting incorrect solutions may enhance sense-making processes (Koedinger et al., 2013).

A second common implementation uses analogous examples. These types of sense-making support typically provide step-by-step guidance for students to map corresponding features across examples so as to extract their commonalities (e.g., Bodemer & Faust, 2006; Gutwill, Frederiksen, & White, 1999; Özgün-Koca, 2008). For example, Gutwill and colleagues (1999) found that asking students to map features of corresponding GRs to one another was effective in enhancing learning outcomes. Providing analogous examples aligns with KLIs worked example's principle (Koedinger et al., 2013).

Studies that compared the effects of sense-making support with linked representations and analogous examples yield contradictory findings. There is evidence in favor of linked representations (e.g., van der Meij & de Jong, 2006, 2011), but there is also evidence in favor of analogous examples (e.g., Gutwill et al., 1999; Özgün-Koca, 2008). Hence, in the present experiment, we compare sense-making support with linked representations and with analogous examples, while incorporating self-explanation prompts in both.

Supporting nonverbal inductive refinement processes in connection making. KLI proposes that learning of simple knowledge components does not require that students engage in verbally mediated learning processes because there is nothing to explain. Evidence for this claim comes from studies showing that sense-making support is ineffective for simple knowledge components in perceptual learning (Schooler, Ohlsson, & Brooks, 1993; Schooler, Fiore, & Brandimonte, 1997) or grammar learning (Wylie, Koedinger, & Mitamura, 2009). KLI identifies a number of principles to guide the design of instructional activities that enhance nonverbal, implicit, inductive processes (Koedinger et al., 2013). Here, we discuss two principles that apply to perceptual fluency in connection making: immediate feedback and exposure to varied instances.

We note that the majority of connection-making support has focused on supporting sense-making processes rather than inductive processes. However, a new line of research yields a type of intervention that aligns with the KLI principles for inductive processes (Kellman & Massey, 2013; Kellman, Massey, & Son, 2010; Wise, Kubose, Chang, Russell, & Kellman, 2000). Kellman and colleagues developed interventions that provide fluency-building support for several science and mathematics topics (Kellman et al., 2009). These interventions ask students to rapidly classify representations over many short problems. In line with the KLI principle of immediate feedback, students receive correctness feedback on these problems. Further, the problems expose students to systematic variation, often in the form of contrasting cases, so that irrelevant features vary but relevant features remain constant across problems (Kellman & Massey, 2013). Studies in several domains (e.g., Kellman & Massey, 2013) show that fluency-

building support leads to large and lasting gains in perceptual fluency that transfer to new instances and to learning gains on domain knowledge tests. Hence, in the present experiment, we investigate the effectiveness of fluency-building problems designed based on Kellman and colleagues' interventions.

Summary and Research Questions

In summary, KLI leads to the hypotheses we test in this article, illustrated in Figure 2. We test the effects of sense-making problems that support verbally mediated sense-making processes (Figure 2, Path 1) to enhance understanding of connections (Figure 2, Path 5), and of fluency-building problems that support nonverbal inductive processes (Figure 2, Path 2) to enhance perceptual fluency (Figure 2, Path 6). We hypothesize that combining both types of connection-making support will enhance students' learning of fractions knowledge (Figure 2, Paths 7 and 8).

This hypothesis remains untested because research on sense-making support and research fluency-building support are, to date, separate lines of research. In particular, prior research on sense-making support did not assess or manipulate students' perceptual fluency. Notably, most research on sense-making support involved connecting a GR to text-based representations. It seems reasonable to assume that students are fluent in reading (i.e., they have a high level of perceptual fluency in processing text). However, we do not know whether students in these studies had some level of perceptual fluency with the GR, and we do not know whether their level of prior perceptual fluency affected their benefit from sense-making support. Likewise, prior research on fluency-building support typically involved students who had already acquired conceptual understanding of the domain knowledge (e.g., Kellman et al., 2009), which is likely to involve understanding of connections. However, we do not know whether students' prior knowledge affected their benefit from fluency-building support.

We conducted a controlled classroom experiment that tested the following research questions and hypotheses:

Research Question 1: Does connection-making support enhance students' learning gains?

Hypothesis 1.1: Students who receive sense-making problems that support connection making show higher learning gains of fractions knowledge than students who do not.

Hypothesis 1.2: Students who receive fluency-building problems show higher learning gains than students who do not.

Hypothesis 1.3: Students who receive a combination of sense-making and fluency-building problems show higher learning gains than students who receive either alone.

Research Question 2: Are sense-making problems more effective if they include linked GRs or analogous examples?

This question was explorative, so we did not test specific hypotheses.

Research Question 3: Does connection-making support enhance students' benefit from MGRs?

Hypothesis 3.1: Students who work with MGRs *without* connection-making support show higher learning gains than students who work with a single GR.

Hypothesis 3.2: Students who work with MGRs *with* connection-making support show higher learning gains than students who work with a single GR.

Classroom Experiment

Method

Experimental design. We randomly assigned individual students to work with one of several versions of the Fractions Tutor, which differed with respect to the types of connection-making problems they included. Our experiment had a $2 \times 3 + 1$ design, summarized in Table 1. The two experimental factors were two types of connection-making problems: sense-making support and fluency-building support. The sense-making factor varied whether students received sense-making problems with linked representations (SL), sense-making problems with analogous examples (SE), or no sense-making problems. This factor was crossed with the fluency-building support factor, which varied whether students received fluency-building problems (F) or not. Students in the MGR condition received MGRs but no connection-making problems. Students in the single-graphical-representation (SGR) condition received only number lines and no connection-making problems.

Participants. There were 599 4th- and 5th-grade students, aged 9–13 years, from five elementary schools (25 classes) in one school district in Pennsylvania who participated in the experiment. The school district was among the 10% highest ranked in reading and mathematics assessments of 500 Pennsylvania public school districts in the year of 2010/2011, with about 12% of students

Table 1
Overview of Experimental Conditions

Fluency-building support		Sense-making support		Control
	No	Linked representations	Analogous examples	
No	Multiple-graphical-representations (MGR)	Sense-making with linked GRs (SL)	Sense-making with analogous examples (SE)	
Yes	Fluency-building (F)	Sense-making with linked GRs plus fluency-building (SL-F)	Sense-making with analogous examples plus fluency-building (SE-F)	
Control				Single-graphical-representation (SGR)

enrolled in free or reduced-price lunch programs, and 95% of the students being White. The school district volunteered to participate in this research.

Instructional materials: The Fractions Tutor. We conducted the experiment in the context of the Fractions Tutor, an effective intelligent tutoring system designed for use in real classrooms (Rau et al., 2013). The Fractions Tutor supports learning through problem solving while providing immediate feedback and on-demand hints, both related to each problem step. The Fractions Tutor emphasizes conceptual learning by emphasizing principled understanding of fractions as proportions of a unit while students solve problems. The curriculum covers 10 topics (see appendix in online supplemental material, Table 1A), covering about 10 hr of instruction. Students worked individually at their own pace. All conditions received 80 tutor problems: eight problems per topic, for 10 fractions topics. For our experiment, we created different versions of the Fractions Tutor that varied what types of support for connection-making competencies it provides, detailed in the following. Consequently, the problems students encountered in the Fractions Tutor differed by condition, but we equated the number of problem-solving steps across conditions. Pilot-testing established that they took about the same time.

SGR condition. Students in the SGR condition worked on number line problems only, eight per topic.

MGR condition. Students in the MGR condition worked on eight individual-representation problems per topic. These problems involved only one GR per problem, but MGRs were used across problems, such that the students encountered each GR an equal number of times. Thus, the MGR condition received all three GRs, but no connection-making problems.

Figure 3 shows an example of an individual-representation problem. As students work through the steps of the problem, the Fractions Tutor provides feedback. The items shown in green are student entries with tutor feedback indicating that the value is correct, such as values entered in input boxes, selections from menus, and dots placed on an interactive number line. Students can also request a hint from the tutor on every step by clicking the brown button at the top right. Students interact with the GRs by using buttons to partition the GR into sections and by clicking to highlight sections in circles and rectangles or to place a dot on the number line. They also receive feedback on these interactions.

In the remaining five conditions, the first four problems for each topic were individual-representation problems. Students received the same number of individual-representation problems for each GR. The last four problems per topic were connection-making problems (i.e., sense-making problems with linked representations, sense-making problems with analogous examples, and/or fluency-building problems), corresponding to the student's experimental condition. Table 2 illustrates how sense-making problems and fluency-building problems were combined by contrasting three of the conditions.

Sense-making with analogous examples (SE) condition. Students in the SE condition received four problems per problem in which they solved a part of the problem with one GR while being able to reference a set of worked-out steps for an analogous example that involved a different GR. These problems all share the

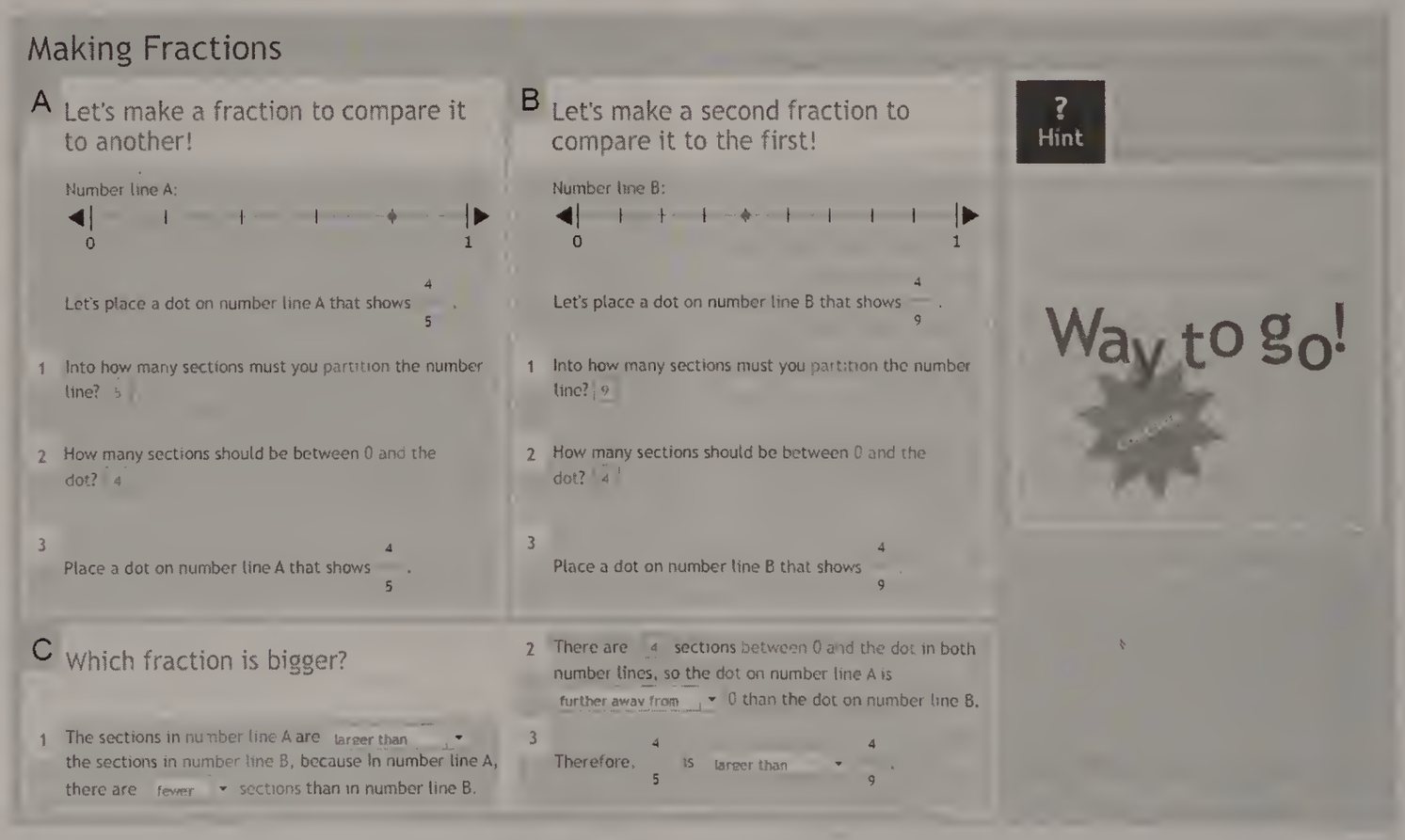


Figure 3. Example of a tutor problem with only the number-line representation. See the online article for the color version of this figure.

Table 2

Problem Sequence Per Condition: For Each Topic, Problems 1–4 (P1–P4) Are Individual-Representation Problems (I); Problems 5–8 Are Connection-Making Problems: Sense-Making Problems With Analogous Examples (SE, Underlined) or Perceptual Fluency-Building Problems (F, Italicized)

Condition	Topic	P1	P2	P3	P4	P5	P6	P7	P8
SE	1	I	I	I	I	*SE*	*SE*	*SE*	*SE*
	2	I	I	I	I	*SE*	*SE*	*SE*	*SE*
	...					...			
F	1	I	I	I	I	*F*	*F*	*F*	*F*
	2	I	I	I	I	*F*	*F*	*F*	*F*
	...					...			
SE-F	1	I	I	I	I	*SE*	*SE*	*F*	*F*
	2	I	I	I	I	*SE*	*SE*	*F*	*F*
	...					...			

Note. Bold-underlined problems and bold-italicized problems are used in the causal path analysis.

same format, illustrated in Figure 4. Students were first given worked-out steps for a question with an area model (i.e., circle or rectangle; Figure 4A, light green panel on the left). Next, the problem-solving part appeared on the right (Figure 4B, light blue panel in the middle), with steps that were analogous to those in the example part. The problem-solving part always involved the number line. The key idea was that the analogous example uses the GR that is more familiar to students, given that—as mentioned above—fractions curricula tend to introduce fractions with area models. After completing the problem, students received self-explanation prompts to abstract a general principle from the two GRs (e.g., that both show equivalent fractions by repartitioning the same amount; Figure 4C, bottom). Self-explanation prompts were implemented in a fill-in-the blank format with drop-down menus on which students receive feedback. Similarly simple formats have been shown to be effective in prior research with intelligent tutoring systems or other educational technologies (Aleven & Koedinger, 2002; Atkinson, Renkl, & Merrill, 2003) and more effective than open-ended forms of self-explanation prompts (Gadgil, Nokes-Malach, & Chi, 2012; Johnson & Mayer, 2010; van der Meij & de Jong, 2011).

Sense-making with linked representations (SL) condition. Students in the SL condition received four problems per topic that included *support to make sense of connections with linked GRs*

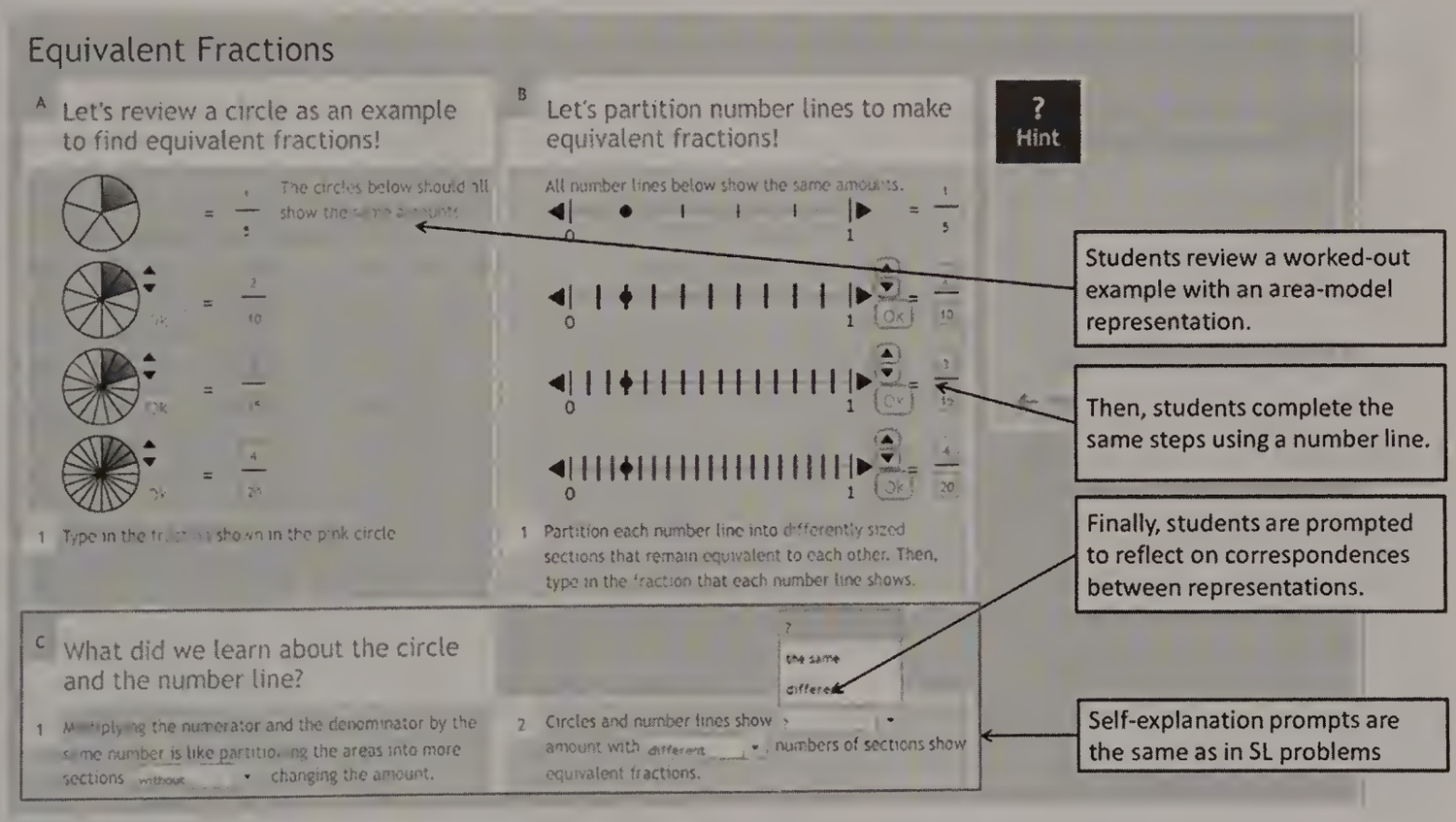


Figure 4. Example of a sense-making problem with analogous examples. The self-explanation prompts in Part C (highlighted in pink) were identical to sense-making problems with linked representations. See the online article for the color version of this figure.

(see Figure 5). Students interacted with a number line (Figure 5A) to solve a problem, while an area model (i.e., a circle or a rectangle) updated automatically to mimic the same steps. Because students tend to be more familiar with area models than with number lines, linking was implemented such that the more familiar GR provided feedback on interactions with the less familiar GR. The SL problems included the same self-explanation prompts as SE problems (Figure 5B).

Fluency-building (F) condition. Students in the F condition received four problems per topic that included *fluency-building support for connection making* (see Figure 6). The fluency-building problems were designed based on Kellman and colleagues' (2010) interventions. Hence, they provided students with numerous short categorization problems. In the equivalent fractions topic, for instance, students sorted a variety of GRs using drag-and-drop (see Figure 6). In alignment with Kellman and colleagues' interventions, fluency-building problems provided only correctness feedback. Students could request hints, but hint messages only provided general encouragement (e.g., "give it a try!"). Finally, the fluency-building problems encouraged visual problem-solving strategies. For example, in the equivalent fractions topic, students were instructed to visually judge equivalence rather than counting sections. To discourage counting strategies, we included examples with sections too small to count.

Combined sense-making and fluency-building conditions. Students in the sense-making with linked representations plus fluency-building (SL-F) condition also received four connection-making problems per topic: two SL problems followed by two F problems. Similarly, students in the sense-making with analogous examples plus fluency-building (SE-F) condition received two SE problems followed by two F problems. We decided to provide sense-making problems before fluency-building problems in each topic because understanding is expected before fluency in educational practice guides (e.g., National Council of Teachers of Mathematics, 2000, 2006).

Test instruments. Students took the tests three times: before they started working with the tutor (pretest), immediately after they finished working with the tutor (immediate posttest), and 1 week after the immediate posttest (delayed posttest). The delayed posttest was included so as to test whether students' knowledge is robust in that it lasts over time (Koedinger et al., 2012). We created three equivalent test forms, which included the same type of problems but with different numbers. We counterbalanced the order in which the different test forms were administered.

The tests targeted robust knowledge of fractions (i.e., with respect to domain knowledge, not connection-making knowledge) considering two knowledge types: procedural and conceptual knowledge. The conceptual scale included eight items that assessed students' principled understanding of fractions. The test items asked students to reconstruct the unit of a fraction, identify fractions from GRs, answer proportional reasoning questions, and complete written reasoning questions about fraction comparison tasks. The procedural scale included nine items that assessed students' ability to solve questions by applying algorithms. The test items asked students to find a fraction between two given fractions using GRs, finding equivalent fractions, addition, and

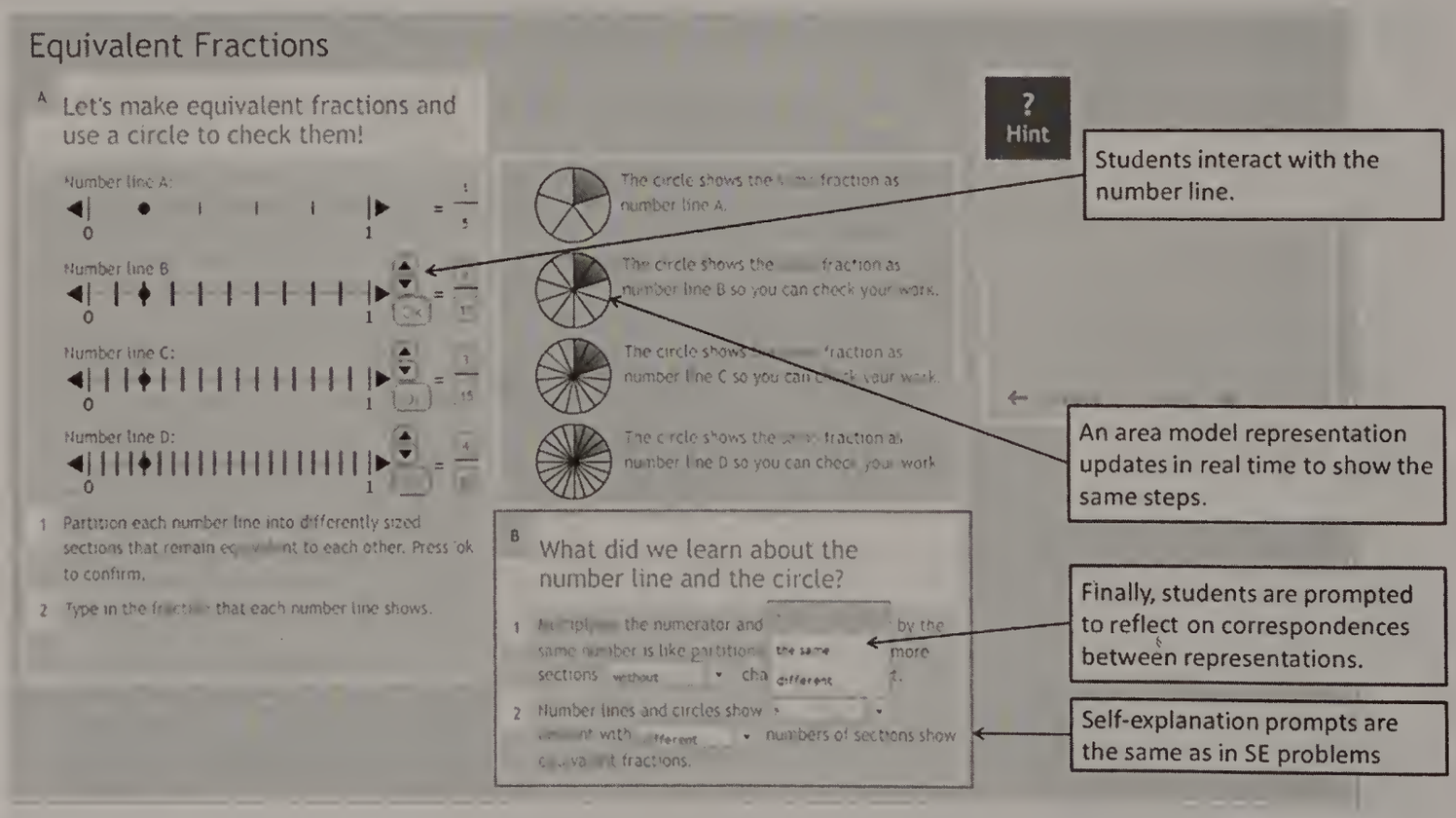


Figure 5. Example of sense-making problem with linked representations. The self-explanation prompts in Part B (highlighted in pink) were identical to sense-making problems with analogous examples. See the online article for the color version of this figure.

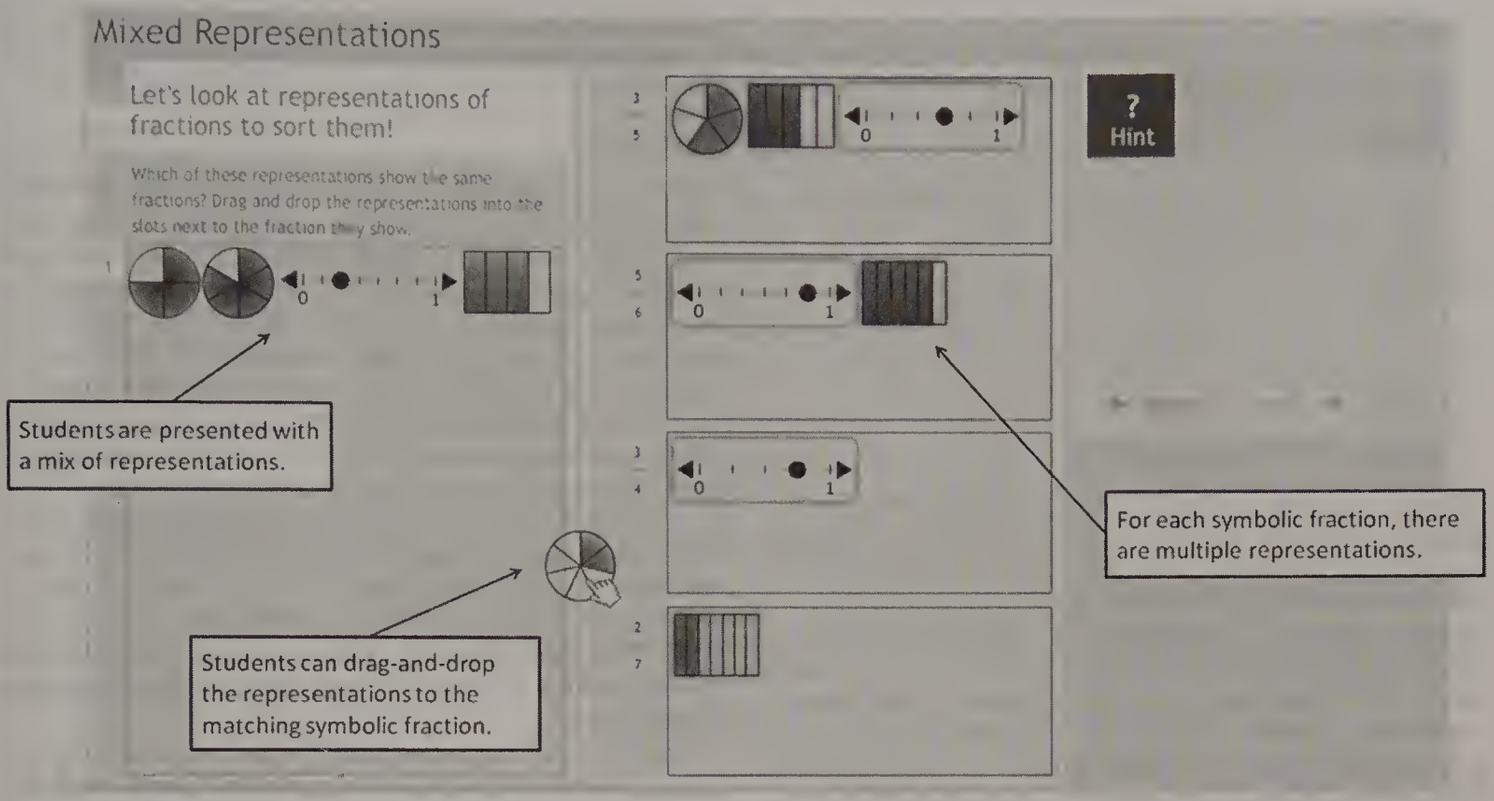


Figure 6. Example of a fluency-building problem. See the online article for the color version of this figure.

subtraction. Both scales included multiple-choice and open-ended items. Half of the items in both test scales were reproduction and transfer items, respectively. Reproduction items were similar to individual-representation problems students had encountered during their work on the tutor. Transfer items were new relative to those covered in the tutor. The goal in including transfer items was to assess whether students' knowledge is robust in that it is transferred to unfamiliar problems (Barnett & Ceci, 2002). Example items for both tests can be found in the appendix in online supplemental material (Figures 1A and 2A). For questions that required multiple steps, partial credit was given for each correct step. The scores reported here are relative scores (i.e., ranging from 0 to 1). The theoretical structure of the test was based on a factor analysis with pretest data from the current experiment and was replicated with data from the immediate and delayed posttests. All test items were evaluated for their difficulty levels and discriminatory power using item-response-theory models. Taken together, the test items covered a range of difficulty levels. All items had good discriminatory power. Both scales had good reliability with Cronbach's α of .70 for the conceptual scale and Cronbach's α of .77 for the procedural scale.

Procedure. The study took place at the beginning of the 2011/2012 school year. Students accessed all materials online from their school's computer lab. They were instructed to work individually at their own pace with the Fractions Tutor. Classroom teachers led the sessions as they normally would during computer-lab hours; that is, they walked around to help individual students who needed assistance. They managed their classrooms in regular fashion; for instance, they told students to be quiet when they were chatting. Experimenters were present for the first 2 days of the experiment to ensure that the Fractions Tutor worked smoothly in the labs.

On Day 1 of the study, students completed a 30-min pretest. They then worked on the Fractions Tutor for about 1 hr per day for 10 consecutive school days (i.e., 2 weeks, yielding about 10 hr spent on the Fractions Tutor in total). On the last day, students completed a 30-min posttest. One week later, students took a delayed posttest.

Analysis. Data in education research often has complex patterns of variance because of to the fact that students are nested within classes (i.e., classes may account for a portion of the variance) and within schools (i.e., schools may account for a portion of the variance). Taking these sources of variance into account in statistical analyses allows to reduce the error variance statistical significance tests (Raudenbush & Bryk, 2002). Hierarchical linear models are a type of statistical model that allows accounting for such nested sources of variance (HLM; Raudenbush & Bryk, 2002).

We tested a number of variables, including teacher, school district, test form sequence, grade level, number of problems completed, total time spent with the tutor, random intercepts and slopes for classes and schools. We also tested whether including each level of the HLM increased model fit. The outcome of this selection procedure was the following four-level HLM. At level 1, we modeled performance on each of the tests for each student. At Level 2, we accounted for differences between students. Level 3 models random differences between classes, and Level 4 random differences between schools. Specifically, we used the following HLM:

$$Y_{ijkl} = (((\mu + W_1) + V_{kl}) + \beta_2 * s_j + \beta_3 * f_j + \beta_4 * p_j + \beta_5 * s_j * p_j + \beta_6 * f_j * p_j + U_{jkl}) + \beta_1 * t_i + R_{ijkl}$$

with

(level 1)

$$Y_{ijkl} = \varepsilon_{jkl} + \beta_1 * t_i + R_{ijkl}$$

(level 2)

$$\varepsilon_{jkl} = \delta_{kl} + \beta_2 * se_j + \beta_3 * sl_j + \beta_4 * f_j + \beta_5 * p_j + \beta_6 * se_j * p_j + \beta_7 * sl_j * p_j + \beta_8 * f_j * p_j + U_{jkl}$$

(level 3)

$$\delta_{kl} = \gamma_1 + V_{kl}$$

(level 4)

$$\gamma_1 = \mu + W_1$$

Table 3 provides an overview of the variables included in the HLM. Index *i* stands for test time (i.e., immediate and delayed posttest), *j* for the student, *k* for class, and *l* for the school. The dependent variable Y_{ijkl} is student$_i$'s score on the dependent measures at test time t_i (i.e., immediate or delayed posttest), ε_{jkl} is the parameter for the intercept for student$_j$'s score, β_1 is the parameter for the effect of test time t_i, β_2 is the parameter for the effect of sense-making problems with analogous examples se_j, β_3 is the parameter for the effect of sense-making problems with linked representations sl_j, β_4 is the parameter for the effect of fluency-building problems f_j, β_5 is the parameter for the effect of student$_j$'s performance on the pretest p_j, β_6 is the parameter for an aptitude-treatment interaction between sense-making problems with analogous examples se_j and student$_j$'s performance on pretest p_j, β_7 is the parameter for an aptitude-treatment interaction between sense-making problems with linked representations sl_j and student$_j$'s performance on pretest p_j, β_8 is the parameter for an aptitude-treatment interaction between fluency-building problems f_j and student$_j$'s performance on pretest p_j, δ_{kl} is the parameter for the random intercept for class$_k$, γ_1 is the parameter for the random intercept for school$_l$, and μ is the overall average. We ran this model in the SAS software package for mixed models.

Table 3
Overview of Variables Included in the HLM

Variable	Explanation
Y_{ijkl}	Student$_i$'s score on the dependent measures at test time t_i (i.e., immediate or delayed posttest)
ε_{jkl}	Intercept for student$_j$'s score
β_1	Effect of test time t_i
β_2	Effect of sense-making problems with prompts for analogical comparisons GRs se_j
β_3	Effect of sense-making problems with linked GRs sl_j
β_4	Effect of perceptual fluency-building problems f_j
β_5	Effect of student$_j$'s performance on the pretest p_j
β_6	Aptitude-treatment interaction between sense-making problems with analogical comparisons se_j and student$_j$'s performance on pretest p_j
β_7	Aptitude-treatment interaction between sense-making problems with linked GRs sl_j and student$_j$'s performance on pretest p_j
β_8	Aptitude-treatment interaction between perceptual fluency-building problems f_j and student$_j$'s performance on pretest p_j
δ_{kl}	Random intercept for class$_k$
γ_l	Random intercept for school$_l$
μ	Overall average

Results

We excluded students who did not complete all tests or did not complete the Fractions Tutor in the time allocated by their classroom teacher because they did not receive the full intervention and did not complete all topics that were tested in the posttests. The final sample included a total of $N = 428$ ($n = 61$ in the SGR condition, $n = 64$ in the MRG condition, $n = 52$ in the SL condition, $n = 59$ in the SE condition, $n = 73$ in the F condition, $n = 61$ in the SL-F condition, $n = 59$ in the SE-F condition). The number of students who were excluded from the analysis did not differ significantly between conditions, $\chi^2(6, N = 169) = 4.34$, $p > .10$. Excluded students had significantly lower pretest scores on the conceptual knowledge test, $F(1, 594) = 6.73$, $p < .05$, and on the procedural knowledge test, $F(1, 594) = 5.60$, $p < .05$, but there were no differences between conditions (Fs < 1). Students' lower prior knowledge may explain why they took longer in working with the Fractions Tutor and, hence, did not finish in the allocated time.

Table 4 shows the means and *SD*s for the conceptual and procedural scales by test time and condition. Table 5 shows the total amount of time spent on tutor problems by condition. To verify that time spent did not differ between conditions, we used the same HLM as described above. There were no significant effects of sense-making support, fluency-building support, nor a significant interaction among these factors on time spent (Fs < 1).

Learning gains. In learning experiments in real educational settings, any difference between conditions needs to be interpreted relative to pretest-to-posttest learning gains (Lipsey et al., 2012). Thus, we first verified whether students learned from the Fractions Tutor. To do so, we used a modified version of the HLM described above on all seven conditions, using pretest scores as a repeated, dependent measure rather than as a covariate (the SAS-code can be found in the appendix in online supplemental material, Figure 3A). Students performed significantly better on conceptual knowledge at the immediate posttest ($p < .0001$, $d = .40$), and at the delayed posttest ($p < .0001$, $d = .60$), compared with the pretest. Students performed significantly better on procedural knowledge at the immediate ($p < .0001$, $d = .20$) and at the delayed posttest ($p < .0001$, $d = .24$), compared with the pretest.

Effects of connection-making support. To investigate Research Question 1, whether a combination of sense-making problems and fluency-building problems leads to higher learning gains than either type of problem alone, we applied the HLM described above to the 2 × 3 design (i.e., without the SGR condition; the SAS-code can be found in the appendix in online supplemental material, Figure 4A). The parameter estimates can be found in the appendix in online supplemental material (Tables 2A for random intercepts, 3A for fixed effects in the conceptual knowledge model, Table 4A for fixed effects in the procedural knowledge model). There were no main effects of sense-making problems (Hypothesis 1.1) or fluency-building problems (Hypothesis 1.2) on conceptual knowledge or on procedural knowledge (Fs < 1). There were no significant interactions of sense-making problems or fluency-building problems with pretest performance. There was no significant interaction on procedural knowledge (Fs < 1).

Table 4
Means (and SDs) for Conceptual and Procedural Knowledge at Pretest, Immediate Posttest, Delayed Posttest

Measure	Condition	Pretest	Immediate posttest	Delayed posttest
Conceptual knowledge	Multiple-graphical-representations (MGR)	.33 (.20)	.45 (.23)	.48 (.26)
	Sense-making with linked GRs (SL)	.38 (.20)	.49 (.23)	.51 (.26)
	Sense-making with analogous examples (SE)	.36 (.22)	.43 (.20)	.49 (.26)
	Fluency-building (F)	.31 (.21)	.37 (.22)	.44 (.24)
	Sense-making with linked representations plus fluency-building problems (SL-F)	.36 (.20)	.43 (.24)	.49 (.25)
	Sense-making with analogous examples plus fluency-building problems (SE-F)	.39 (.21)	.52 (.24)	.58 (.26)
	Single-graphical-representation (SGR)	.37 (.20)	.43 (.25)	.48 (.20)
Procedural knowledge	Multiple-graphical-representations (MGR)	.25 (.25)	.30 (.28)	.30 (.26)
	Sense-making with linked representations (SL)	.21 (.18)	.26 (.24)	.26 (.24)
	Sense-making with analogous examples (SE)	.26 (.21)	.29 (.24)	.31 (.27)
	Fluency-building condition (F)	.19 (.17)	.23 (.20)	.25 (.22)
	Sense-making with linked representations plus fluency-building problems (SL-F)	.20 (.18)	.25 (.21)	.26 (.21)
	Sense-making with analogous examples plus fluency-building problems (SE-F)	.26 (.20)	.32 (.26)	.33 (.26)
	Single-graphical-representation (SGR)	.21 (.20)	.25 (.22)	.27 (.23)

Note. Min. score is 0, max. score is 1.

However, there was a significant interaction between sense-making problems and fluency-building problems on conceptual knowledge, $F(2, 343) = 4.11$, $p = .017$, $\eta^2 = .03$, such that students who received both types of problems performed best on the conceptual posttests. To gain further insights into this interaction effect, we turn to Research Question 2: are sense-making problems more effective if they include linked GRs or analogous examples? We examined simple effects of the sense-making factor for the conditions with fluency-building problems (i.e., SL-F, SE-F, and F conditions) and without fluency-building problems (i.e., SL, SE, and MGR conditions). On conceptual knowledge, there was a significant effect of sense-making problems among the conditions with fluency-building problems, $F(2, 343) = 4.34$, $p = .014$, $\eta^2 = .07$, such that the SE-F condition significantly outperformed the F condition, $t(341) = 2.82$, $p = .005$, $d = .32$, and the SL-F condition, $t(342) = 2.20$, $p = .05$, $d = .26$. The difference between the SE-F condition and the F condition was not significant ($t < 1$). The effect of sense-making problems was not significant for the conditions without fluency-building problems ($F < 1$), and consequently, none of the post hoc comparisons were significant.

Table 5
Means (and SDs) of Total Time Spent on Tutor Problems by Condition

Condition	Time on tutor in minutes
Multiple-graphical-representations (MGR)	232.04 (62.88)
Sense-making with linked GRs (SL)	206.27 (60.3)
Sense-making with analogous examples (SE)	213.7 (58.32)
Fluency-building (F)	199.25 (54.97)
Sense-making with linked representations plus fluency-building problems (SL-F)	215.83 (58.43)
Sense-making with analogous examples plus fluency-building problems (SE-F)	203.51 (53.61)
Single-graphical-representation (SGR)	189.47 (41.54)

To investigate whether MGRs are more effective than an SGR (Hypotheses 3.1 and 3.2), we applied a modified version of the HLM described above to the SGR, MGR, and SE-F condition (i.e., the most successful connection-making condition; the SAS-code can be found in the appendix in online supplemental material, Figure 4A). There were no significant differences between the MGR condition and the SGR condition ($ps > .10$; Hypothesis 3.1). The SE-F condition significantly outperformed the SGR condition on conceptual knowledge, $t(115) = 2.41$, $p = .016$, $d = .27$, but not on procedural knowledge ($t < 1$; Hypothesis 3.2).

Discussion

With respect to Research Question 1 (does connection-making support enhance students' learning gains?), our results do not support Hypotheses 1.1 or 1.2, that problems that work on sense-making or working on fluency-building problems would enhance robust fractions knowledge, respectively. However, our results support Hypothesis 1.3 for conceptual knowledge: working on a combination of sense-making problems and fluency-building problems was effective. Somewhat to our surprise, neither type of connection-making support alone, but *only* the combination of both was effective. With respect to Research Question 3 (Does connection-making support enhance students' benefit from MGRs?), our results stand in contrast to Hypothesis 3.1 but support Hypothesis 3.2. Comparisons to the SGR condition show that students did not benefit from working with MGRs, unless they received a *combination* of sense-making and fluency-building support.

We did not find significant effects on procedural knowledge. It may be that students' conceptual knowledge benefits from connection making because each representation provides a different conceptual view on what fractions are, whereas procedural knowledge may rely more on experience with algorithmic operations tasks rather than on conceptual understanding.

With respect to our exploratory Research Question 2, whether problems that help students make sense of connections are more

effective if they include linked GRs or analogous examples, we find that analogous examples lead to higher learning gains on a test of robust fractions knowledge than linked GRs.

Causal Path Analysis Modeling

The experiment showed that only the *combination* of sense-making problems and fluency-building problems was effective in enhancing students' learning of domain knowledge. This finding leads to open questions about *how* sense-making processes and inductive refinement processes interact (Figure 2, Paths 3 and 4). Hence, we seek to better understand the nature of this interaction through an additional data source—the tutor log data as an indicator of problem-solving performance—using causal path analysis modeling. The logs provide a detailed record of students' interactions with the Fractions Tutor at the "transaction" level (i.e., attempts at steps, hint requests, etc.). Given that sense-making problems with analogous examples were more effective than those with linked GRs, we focused on the SE conditions in this analysis.

Hypotheses

We investigate two possible mechanisms by which sense-making problems and fluency-building problems might interact. One mechanism may be that working on fluency-building problems enhances students' benefit from sense-making problems (Figure 2, Path 3; we will refer to this as the *fluency hypothesis*). According to the *fluency hypothesis*, perceptually fluent students may benefit from increased cognitive capacity during subsequent learning tasks (Kellman et al., 2009; Koedinger et al., 2012). Therefore, they should show higher performance on sense-making problems. We contrast the fluency hypothesis to the *practice hypothesis* that receiving more practice on sense-making problems leads to higher performance on sense-making problems. The SE condition provides four sense-making problems per topic, whereas the SE-F condition provides only two sense-making problems per topic. Therefore, the practice hypothesis predicts that the SE condition should show higher performance on sense-making problems than the SE-F condition. To see the effect of having practice with fluency-building problems on students' performance on sense-making problems, we compare the SE condition to the SE-F condition. In the SE condition, problems P5, P6, P7, and P8 were sense-making problems (for each of the 10 topics, see Table 2). In the SE-F condition, only problems P5 and P6 were sense-making problems (for each of the 10 topics). Hence, when comparing the SE and SE-F conditions, problems P5 and P6 of each topic serve as the basis for the comparison (bold-underlined problems in Table 2).

Another mechanism may be that working on sense-making problems enhances students' benefit from fluency-building problems (Figure 2, Path 4; *sense-making hypothesis*). Prior research shows that students have difficulties in making sense of connections at a conceptual level and typically do not make connections spontaneously (Ainsworth et al., 2002; Rau et al., 2014). Therefore, the sense-making hypothesis predicts that students may not be able to discover what features of the GRs depict meaningful information while working on fluency-building problems, which may lead to inefficient learning strategies (e.g., trial-and-error) that can impede their benefit from fluency-building problems. In particular, the visual features that denote fractions may not be easy to detect, and can perhaps not be learned in a purely inductive manner. Therefore, sense-making support could increase students' performance on fluency-building problems. We contrast the sense-making hypothesis to the practice hypothesis that receiving more practice on fluency-building problems leads to higher performance on fluency-building problems. The F condition provides four fluency-building problems per topic, whereas the SE-F condition provides two fluency-building problems per topic. Therefore, the practice hypothesis predicts that the F condition should show higher performance on fluency-building problems than the SE-F condition. To investigate the effect of having practiced on sense-making problems on students' performance on fluency-building problems, we compare the F condition to the SE-F condition. In the F condition, problems P5, P6, P7, and P8 were fluency-building problems (for each of the 10 topics, see Table 2). In the SE-F condition, only problems P7 and P8 were sense-making problems (for each of the 10 topics). Hence, when comparing the F and SE-F conditions, problems P7 and P8 for each topic serve as the basis for the comparison (bold-italicized problems in Table 2). In testing the fluency hypothesis and the sense-making hypothesis, we allow for the possibility that they are not mutually exclusive.

Method

To investigate these hypotheses, we use causal path analysis, which provides a unified framework to test mediation hypotheses, estimate total effects, and separate direct from indirect effects in a coherent statistical model (Bollen & Pearl, 2013; Chickering, 2002; Spirtes et al., 2000). We constructed causal path analysis models that correspond to the fluency hypothesis and to the sense-making hypothesis, respectively.

Because we selected the SE and SE-F conditions for the fluency hypothesis model and the F and SE-F conditions for the sense-making hypothesis model, 190 students were included in the analysis (n = 59 in the SE condition, n = 73 in the F condition, and n = 58 in the SE-F condition).We operationalized performance on the tutor problems as error rates: making fewer errors while solving a tutor problem indicates higher problem-solving performance. Rather than using the overall error rate, we classified errors based on the detailed knowledge components to which they relate. For the fluency hypothesis model, we computed the error rate for each knowledge component across the sense-making problems P5 and P6 for all 10 topics (bold-underlined problems in Table 2). For the sense-making hypothesis model, we computed the error rate for each knowledge component across the fluency-building problems P7 and P8 for all 10 topics (bold-italicized problems in Table 2). Altogether, the knowledge component model yielded 12 error types that students could make on sense-making problems, and 11 error types that students could make on fluency-building problems, summarized in Tables 6 and 7. Next, included only those error types in the causal path analysis model that (a) were significant predictors of performance on the conceptual posttest, while controlling for pretest, and (b) significantly differed between conditions (i.e., the italicized error types in Tables 6 and 7).

We constructed the causal path analysis models using an automatic algorithm that searches for models that are theoretically plausible and consistent with the data; namely, the Tetrad IV

Table 6
Error Types on Fluency-Building Problems and Number of Occurrences Per Condition (Summed Up for All Students Across Fluency-Building Problems P7 and P8)

Error type	Knowledge component	Number in F	Number in SE-F
nameCircleMixed-Error	Finding circle representations that show the same fraction as a number line or a rectangle	355	126
equivalenceError	Finding equivalent fraction representations	2,899	2,157
improperMixed-Error	Finding representations of improper fractions	1,380	1,608
additionMixedError	Finding representations that show the addend of a given sum equation depicted by representations	207	176
compareMixed-Error	Finding representations that show a fraction smaller or larger than the given one	436	307
diffMixedError	Finding representations that show the difference of two fractions	282	238
nameNLMixed-Error	Finding number line representations that show the same fraction as a circle or a rectangle	949	599
nameRectMixed-Error	Finding rectangle representations that show the same fraction as a number line or a circle	385	133
subtractionMixed-Error	Finding representations that show the subtrahend of a given difference equation depicted by representations	214	240
sumMixedError	Finding representations that show the sum of two fractions	256	205
unitMixedError	Finding the unit of a given fraction	1,050	1,138

Note. Italicized error types were selected for further analysis.

program's[1] GES algorithm. Tetrad IV allows us to specify assumptions that constrain the space of models searched (Chickering, 2002; Spirtes et al., 2000) and to find the model with the best model fit among models that are theoretically tenable and compatible with the experimental design (Spirtes et al., 2000). *Independent variables* in the causal path analysis were sense-making support and fluency-building support. *Dependent variables* were students' performance on the conceptual pretest, immediate, and delayed posttest. *Mediators* were error types students made on the sense-making problems for the fluency hypothesis model, and error types students made on the fluency-building problems in the sense-making hypothesis model.

When conducting a model search, we can narrow the search space based on the knowledge we have about the nature of our data (Spirtes et al., 2000). We assumed that the experimental conditions are exogenous and causally independent, that the pretest was not influenced by the conditions, that the pretest is an exogenous variable and causally independent of the conditions. Furthermore, we assume that the mediators are before the immediate posttest and the delayed posttest, and that the immediate posttest is before the delayed posttest. The search space is defined by the fully saturated model for each hypothesis because it contains all possible edges (or "effects") compatible with these assumptions and with the experimental design.

We had Tetrad search among models that had all, none, or a subset of the edges in the fully saturated model. In the model search, each edge is automatically evaluated as to whether including it yields a better model fit than not including it, and whether it represents a statistically significant effect. Figure 7 (left) illustrates the fully saturated model for the fluency hypothesis (that includes only performance variables related to sense making as possible mediators). Figure 7 (right) illustrates the fully saturated model for the sense-making hypothesis (that includes only performance variable related to perceptual fluency as possible mediators). Thus, Figure 7 illustrates that, even with our assumptions, the search space contains at least 2^{15} (over 32 thousand) distinct path models that are plausible tests for the sense-making hypothesis, and 2^{20} (over 1 million) for the fluency hypothesis. The outcomes of the model search are two causal path analysis models, one corresponding to the fluency hypothesis, one corresponding to the sense-making hypothesis, each consistent with the data and hence allowing us to trust the parameters of the model.

Results

To test the fluency hypothesis, we inspect the model shown in Figure 8, which is the best-fitting model Tetrad IV found for the fluency hypothesis. The model fits the data well ($\chi^2 = 8.32$, $df = 5$, $p = .14$; comparative fit index [CFI] = 0.9943; root mean square error of approximation [RMSEA] = 0.0808).[2] The standardized coefficients and their standard errors, the significance tests for each effect, and the implied covariance matrices for the model are provided in the appendix in online supplemental material (Tables 5A, 6A, and 7A). Figure 8 shows unstandardized coefficients, which are easier to interpret with respect to the effects of number of errors students made. Further, because scores on all tests range between 0 and 1, the effects on the posttests are easy to compare even though coefficients are unstandardized. Recall that this model compares the SE and SE-F conditions based on errors

[1] Tetrad, freely available at www.phil.cmu.edu/projects/tetrad, contains a causal model simulator, estimator, and over 20 model search algorithms, many of which are described and proved asymptotically reliable in (Spirtes, Glymour, & Scheines, 2000).

[2] The usual logic of hypothesis testing is inverted in path analysis. The p-value reflects the probability of seeing as much or more deviation between the covariance matrix implied by the estimated model and the observed covariance matrix, conditional on the null hypothesis that the model that we estimated was the true model. Thus, a low p-value means the *model* can be rejected, and a high p-value means it cannot. Conventional thresholds are .05 or .01, but like other α values, this is somewhat arbitrary. The p-value should be higher at low sample sizes and lowered as the sample size increases, but the rate is a function of several factors, and generally unknown.

Table 7
Error Types on Sense-Making Problems and Number of Occurrences Per Condition (Summed Up for All Students Across Sense-Making Problems P5 and P6)

Error type	Knowledge component	Number in SE	Number in SE-F
place1Error	Locating 1 on the number line given a dot on the number line and the fraction it shows	150	222
selfExplanationError	Incorrect response to self-explanation prompt	1,320	1,629
comparisonError	Comparing two fractions	92	82
denomError	Entering the denominator of a fraction	972	837
equivalence-CompareError	Judging whether two fractions are equivalent	19	18
multiplyError	Entering a number by which to multiply numerator or denominator to expand a given fraction	30	29
nlPartitionError	Partitioning the number line to show an equivalent fraction	1,913	2,115
numberSections-UnitError	Finding the denominator of a fraction by indicating how many sections the unit was divided into	41	44
numError	Entering the numerator of a fraction	1,559	1,390
placeDotError	Placing a dot on the number line to show a fraction	198	253
sectionsBetween-0–1	Indicating that the denominator in a number line is shown by the sections between 0 and 1	61	44
unitError	Selecting the unit for a fraction given the symbolic fraction and a graphical representation	123	115

Note. Italicized error types were selected for further analysis.

students made on the sense-making problems. Further recall that, according to the fluency hypothesis, we expect that practice on fluency-building problems reduces error rates on sense-making problems, and that error rates on sense-making problems mediates the effect of condition on the posttests. Finally, recall that the alternative practice hypothesis suggests that, because students in the SE-F condition have less practice on sense-making problems, they should show higher rates of sense-making errors. The model in Figure 8 shows that students in the SE-F condition, compared to the SE condition, made *more* selfExplanationErrors (i.e., the average student in the SE-F condition made 5.662 more errors in answering self-explanation prompts than the average student in the SE condition, and for each of these errors, the student looses .005 points on the final posttest) and *more* place1Errors (i.e., errors in finding 1 on a number line). Both decreased learning gains. Thus, students' performance on sense-making problems mediated a negative effect of fluency-building support on students' posttest performance. This negative mediation effect is in line with the alternative hypothesis that practice alone explains performance on sense-making problems. In addition, in line with the overall finding of the experiment, Figure 8 shows that fluency-building support had a direct positive effect on posttest performance, which was stronger than the negative mediation effects. That is, the direct path of .116 is larger than the sum of the mediating paths ($-.005 * 5.662 + -.012 * .166 * 5.662$).

To test the sense-making hypothesis, we inspect the model in Figure 9, which shows the best-fitting model for the sense-making hypothesis. This model fits the data reasonably well ($\chi^2 = 16.10$, $df = 6$, $p = .013$; CFI = 0.9822; RMSEA = 0.1338).[3] The standardized coefficients and their standard errors, the significance tests for each effect, and the implied covariance matrices for the model are provided in the appendix in online supplemental material (Tables 5A, 6A, and 7A). Figure 9 shows the unstandardized coefficients. Recall that this model compares the F and SE-F conditions based on errors students made on the fluency-building problems. Further recall that, according to the sense-making hypothesis, we expect that practice on sense-making problems leads to a lower rate of errors on the fluency-building problems, which in turn mediates the effect of condition on the posttests. Finally, recall that the alternative practice hypothesis suggests that, because students in the SE-F condition have less practice on fluency-building problems, they should show higher error rates on fluency-building problems. The model in Figure 9 shows that students in the SE-F condition made more nameCircleMixed errors (i.e., errors in identifying the fraction depicted by a circle) but fewer improperMixedErrors (i.e., errors in identifying an improper fraction) and fewer equivalence errors (i.e., errors in identifying equivalent fractions) than students in the F condition. Students who made fewer nameCircleMixedErrors also made more subtractionMixedErrors (i.e., errors in finding the difference between two given fractions) and improperMixedErrors, which decreased performance in the conceptual posttest. Thus, performance on fluency-building problems mediated the positive effect of sense-making support on the conceptual posttest. There were no additional direct effects of sense-making support on posttest, so that students' higher performance on fluency-building problems fully mediated the positive effect of sense-making support on learning gains.

Discussion

The results from the causal path analysis are consistent with the sense-making hypothesis but stand in contrast to the fluency hypothesis: we did not find evidence that working on fluency-building problems *helps* students benefit from sense-making problems, but that fluency-building problems *decrease* their per-

[3] Ibid. It is worth noting that this model asserts that any effect the SE-F condition (compared to the F condition) has on the post-test or delayed post-test is entirely mediated by the three variables measuring error rates. Thus, it makes a bold and easily falsifiable prediction that is tested by this model.

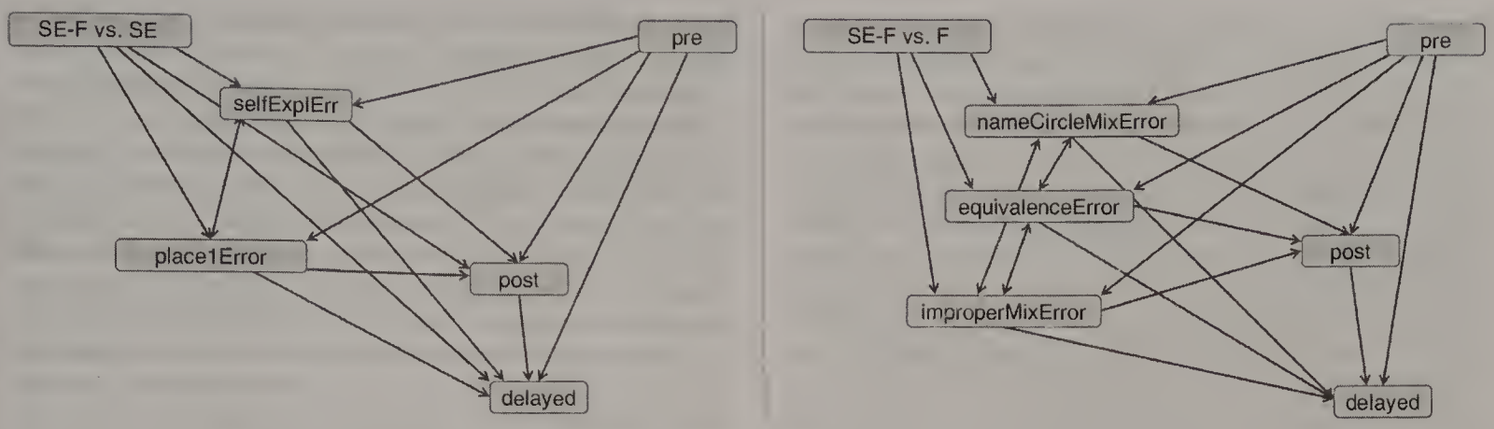


Figure 7. Saturated models for the fluency hypothesis (left) and the sense-making hypothesis (right). See the online article for the color version of this figure.

formance on sense-making problems. Thus, the mediation effect shown in Figure 8 suggests that receiving fluency-building problems comes at the cost of lower performance on sense-making problems: students tend to make more selfExplanationErrors and more place1Errors. Recall that students in the SE condition work on twice as many sense-making problems than students in the SE-F condition, so they receive more practice on these problems compared to the SE-F condition (see Table 2). Hence, the practice hypothesis predicts that they perform somewhat worse on those problems, simply because they have less practice. The model in Figure 8 is in line with the practice hypothesis. Furthermore, the model in Figure 8 puts the performance on sense-making problems in relation to learning gains: higher performance on sense-making problems is associated with higher learning benefit from sense-making problems. However, because we do not find evidence that fluency-building problems help students learn from sense-making problems, our results do not support the fluency hypothesis.

By contrast, the results from the causal path analysis models are in line with the sense-making hypothesis: working on sense-making problems helps students learn from fluency-building problems. The model in Figure 9 demonstrates that, although students who receive sense-making problems make more nameCircleMixedErrors, they make fewer equivalenceErrors and improperMixedErrors while working on fluency-building problems. The reduction of equivalenceErrors and improperMixedErrors mediates the effect of sense-making support on learning gains. NameCircleMixedErrors are confined to an early topic in the Fractions Tutor, whereas equivalenceErrors and improperMixedErrors occur later in the Fractions Tutor. The results, therefore, suggest that working on sense-making problems reduces errors later during the learning phase, which leads to higher learning gains. This finding is particularly interesting because it indicates that having worked on sense-making problems leads to higher performance on fluency-building problems, *even though* students in the F condition had more practice opportunities on fluency-building problems (practice hypothesis). Thus, it seems that sense-making problems prepare students to benefit from subsequent fluency-building problems—even more so than practice with fluency-building problems does.

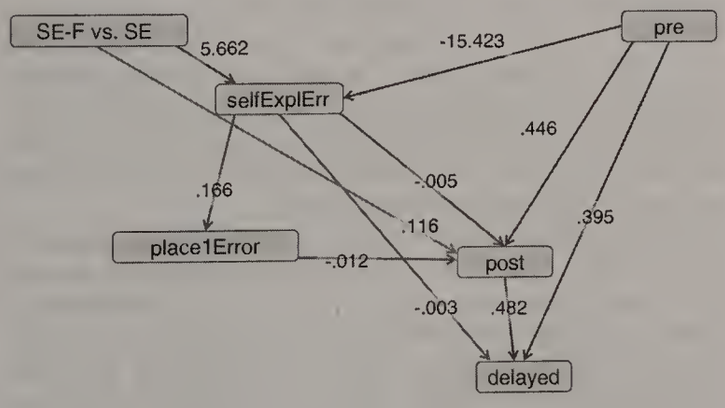


Figure 8. Fluency-hypothesis model with unstandardized parameter estimates. Paths that describe a negative effect of fluency-building support on posttest performance (immediate and final) are highlighted in red, paths that describe a positive effect are highlighted in green. See the online article for the color version of this figure.

General Discussion

Our experiment investigated how best to support students in making connections among MGRs. Our results support our hypothesis that a combination of sense-making and fluency-building support is most effective with respect to learning of conceptual knowledge. Surprisingly, we found that *only* the combination of sense-making problems and fluency-building problems is effec-

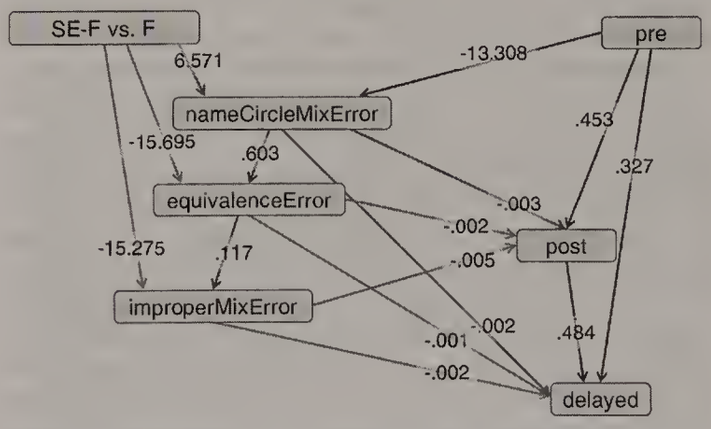


Figure 9. Sense-making hypothesis model with unstandardized parameters. See the online article for the color version of this figure.

tive; taken alone, neither sense-making problems nor fluency-building problems were effective. By establishing that sense-making problems and fluency-building problems interact, this finding extends prior research that has—to the best of our knowledge—exclusively focused on either sense-making support (e.g., Bodemer & Faust, 2006; Seufert, 2003; van der Meij & de Jong, 2006), or on fluency-building support (e.g., Kellman et al., 2009). As argued above, students in prior research on sense-making support may have had some level of perceptual fluency in interpreting the representations used in these studies (i.e., mostly text-based and numerical representations). Likewise, students in prior research on fluency-building support likely had, to some extent, understanding of connections because they had typically received prior instruction on the domain knowledge. Our finding that both types of support are necessary does not necessarily contradict prior research. Rather, our findings extend it by indicating that the aspects that were held constant across conditions in prior research may be an important prerequisite to the effectiveness of either type of support. At a practical level, our results suggest that standard sense-making support should take into account students' level of perceptual fluency. Instructors may need to ensure that students are indeed perceptually fluent in making connections, in which case sense-making support alone could be effective (although this hypothesis has not been tested), or they might need to combine sense-making support with fluency-building support (as in our experiment).

It is also interesting to reflect on the fact that we did not find evidence that MGRs without connection-making support lead to higher learning gains than a single GR that is considered the "superior" GR by some researchers: the number line (National Mathematics Advisory Panel, 2008; Siegler et al., 2010). We found that MGRs were more effective than a single GR *only if* students received a combination of sense-making and fluency-building support. This finding is in line with our own prior research (Rau, Aleven, Rummel, & Rohrbach, 2012), which shows that MGRs are not always effective in enhancing fractions learning. It is also in line with experiments in other domains that failed to show a benefit of MGRs over learning with a single GR (e.g., Berthold & Renkl, 2009; Corradi, Elen, & Clareboug, 2012). MGRs are commonly used in instruction because they emphasize multiple conceptual perspectives. Our results support this practice but also caution that integrating these conceptual perspectives into their domain knowledge is a difficult task for students. To support them in doing so, instruction may need to provide a combination of sense-making support and fluency-building support.

The causal path analysis models provide additional insights into the mechanisms underlying this finding. We found that sense-making problems enhance students' benefit from fluency-building problems by reducing the number of certain types of errors students make on fluency-building problems. Hence, understanding of connections seems to provide the foundation for inductive processes that students engage in when working on fluency-building problems. Our findings do not support the reverse conclusion: we have no evidence that fluency-building problems enhance students' benefit from sense-making problems. In contrast, we found that more practice on sense-making problems yields higher performance on sense-making problems, as expected purely based on practice effects. Thus, it seems that, even if there are benefits of additional cognitive headroom as a result of perceptual fluency, they do not outweigh the advantages of practice effects on the same type of problem.

Lipsey and colleagues (2012) suggest that effect sizes of interventions obtained in real classrooms must be interpreted in relation to pretest-to-posttest changes. Ranging between $d = .20$ and $d = .60$ resulting from a 10-hr long intervention, effect sizes of learning gains are of small to medium size. According to Hattie's (2012) meta-analysis of educational interventions in realistic settings, the average effect size of interventions are $d = .40$ per year on student achievement (e.g., p. 16, p. 240 in Hattie, 2012). Thus, our experiment shows learning gains that compare favorably to those obtained in other studies. A similar argument can be made when interpreting the effect sizes for the between-condition effects. The advantage of receiving a combination of sense-making and fluency-building support compared with working with only the number line representation had an effect size of $d = .27$. Thus, comparing this difference to the learning gain of $d = .40$ on the conceptual knowledge test, the benefit of combining sense-making problems and fluency-building problems when providing students with MGRs seems meaningful.

It is important to note a number of limitations of this research. First, we excluded students who did not finish their work with the Fractions Tutor because they did not receive full exposure to the experimental intervention and because the posttests assessed knowledge targeted in all topics of the curriculum. However, this decision led to excluding many students, and these students had lower pretest scores than students who were included in the analysis. Because students were randomly assigned to conditions and because the number of excluded students did not differ by conditions, this exclusion does not undermine our overall conclusions, but implies that future research should test that our findings generalize to lower-performing students. We also note that the school population was mostly White and included only a small portion of students from low-income families. Although we cannot think of a reason why students from more diverse backgrounds would not benefit from a combination of sense-making and fluency-building support, future research should empirically verify this prediction.

The causal path analysis was limited because (unlike the HLM), it does not allow us to take into account variance because of students being nested in classes and schools. Not taking into account these sources of variance means that the error variance in the causal path analysis is larger than in the HLM analysis, which reduces the statistical power of the analysis. While this limitation does not affect the internal validity of the results, the lower power of the analysis means that there might be effects in the data that we did not detect. Future research should address this issue by using a larger sample for a causal path analysis.

We also note limitations resulting from the presentation of instructional materials. We conducted our experiment in the context of an intelligent tutoring system, an effective type of educational technology that is widely used in classrooms across the United States. Even though this context represents a realistic educational scenario, further research should test whether our results generalize to out-of-technology contexts. For example, future research should investigate whether our findings generalize to contexts in which students use physical representations or a combination of physical and virtual representations. Further, students received sense-making problems before fluency-building

problems. Because this sequence was repeated for each topic of the tutor, we believe that it does not affect the validity of the effects we found in the causal path analysis. However, the effects of fluency-building problems on students' performance on sense-making problems may have been stronger if fluency-building problems had been directly followed by sense-making problems (rather than by individual-representation problems). This limitation may have affected the power of the analysis, but not the validity: we may not have detected all effects, but we can trust the effects that we did detect, and we can trust that the effects we did detect are stronger than the effects we may not have detected.

Conclusions

We tested a prediction that resulted from applying KLI to the case of connection making among MGRs; namely, that students will benefit most from support that targets verbally mediated sense-making processes through which students acquire understanding of connections, and support that targets nonverbal, inductive processes through which students acquire perceptual fluency in making connections. Our experiment extends prior research that has only focused on either sense-making support (e.g., Bodemer & Faust, 2006; Seufert, 2003; van der Meij & de Jong, 2011) or fluency-building support (e.g., Kellman et al., 2009; Kellman & Massey, 2013), but has not investigated potential interactions between these two types of connection-making support. Our results were more pronounced than expected: the combination of sense-making support and fluency-building support was *necessary* for students to benefit from MGRs, compared to a single GR. The causal path analysis suggests sense-making support provides the foundation for students' benefit from fluency-building support. This finding yields a new testable hypothesis: students will learn best when sense-making support is provided before fluency-building support rather than vice versa.

Given the pervasiveness of MGRs in STEM and the well-documented need for connection-making support, our findings have the potential to apply to many domains. The research presented in this article is only a first step in this direction, and we hope it will inspire future research on sense making and perceptual fluency in connection making.

References

Acevedo Nistal, A., Van Dooren, W., & Verschaffel, L. (2013). Students' reported justifications for their representational choices in linear function problems: An interview study. *Educational Studies, 39*(1), 104–117. http://dx.doi.org/10.1080/03055698.2012.674636

Acevedo Nistal, A., Van Dooren, W., & Verschaffel, L. (2015). Improving students' representational flexibility in linear-function problems: An intervention. *Educational Psychology, 34,* 763–786. http://dx.doi.org/10.1080/01443410.2013.785064

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16,* 183–198. http://dx.doi.org/10.1016/j.learninstruc.2006.03.001

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences, 11,* 25–61. http://dx.doi.org/10.1207/S15327809JLS1101_2

Ainsworth, S. E., & van Labeke, N. (2002). *Using a multi-representational design framework to develop and evaluate a dynamic simulation environment.* Paper presented at the International Workshop on Dynamic Visualizations and Learning, Tuebingen, Germany.

Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science: A Multidisciplinary Journal, 26,* 147–179.

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning From Studying Examples to Solving Problems: Effects of Self-Explanation Prompts and Fading Worked-Out Steps. *Journal of Educational Psychology, 95,* 774–783. http://dx.doi.org/10.1037/0022-0663.95.4.774

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128,* 612–637. http://dx.doi.org/10.1037/0033-2909.128.4.612

Behr, M. J., Post, T. R., Harel, G., & Lesh, R. (1993). Rational numbers: Toward a semantic analysis - emphasis on the operator construct. In T. P. Carpenter, E. Fennema, & T. A. Romberg (Eds.), *Rational numbers: An integration of research.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Berthold, K., Eysink, T. H. S., & Renkl, A. (2008). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science, 27,* 345–363.

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Research, 101,* 70–87.

Bodemer, D., & Faust, U. (2006). External and mental referencing of multiple representations. *Computers in Human Behavior, 22,* 27–42. http://dx.doi.org/10.1016/j.chb.2005.01.005

Bodemer, D., Ploetzner, R., Feuerlein, I., & Spada, H. (2004). The Active Integration of Information during Learning with Dynamic and Interactive Visualisations. *Learning and Instruction, 14,* 325–341. http://dx.doi.org/10.1016/j.learninstruc.2004.06.006

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). the Netherlands: Springer. http://dx.doi.org/10.1007/978-94-007-6094-3_15

Charalambous, C. Y., & Pitta-Pantazi, D. (2007). Drawing on a theoretical model to study students' understandings of fractions. *Educational Studies in Mathematics, 64,* 293–316. http://dx.doi.org/10.1007/s10649-006-9036-2

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science: A Multidisciplinary Journal, 13,* 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477. http://dx.doi.org/10.1016/0364-0213(94)90016-7

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research, 3,* 507–554.

Cook, M., Wiebe, E. N., & Carter, G. (2008). The influence of prior knowledge on viewing and interpreting graphics with macroscopic and molecular representations. *Science Education, 92,* 848–867. http://dx.doi.org/10.1002/sce.20262

Corradi, D., Elen, J., & Clareboug, G. (2012). Understanding and enhancing the use of multiple external representations in chemistry education. *Journal of Science Education and Technology, 21,* 780–795. http://dx.doi.org/10.1007/s10956-012-9366-z

Cramer, K. (2001). Using models to build an understanding of functions. *Mathematics Teaching in the Middle School, 6,* 310–318.

Cramer, K., Wyberg, T., & Leavitt, S. (2008). The role of representations in fraction addition and subtraction. *Mathematics Teaching in the Middle School, 13*(8), 490–496.

de Jong, T., Ainsworth, S. E., Dobson, M., Van der Meij, J., Levonen, J., Reimann, P., & Swaak, J. (1998). Acquiring knowledge in science and mathematics: The use of multiple representations in technology-based

learning environments. In M. W. Van Someren, W. Reimers, H. P. A. Boshuizen, & T. de Jong (Eds.), *Learning with multiple representations*. Bingley, United Kingdom: Emerald Group Publishing Limited.

diSessa, A. A., & Sherin, B. L. (2000). Meta-representation: An introduction. *The Journal of Mathematical Behavior, 19,* 385–398. http://dx.doi.org/10.1016/S0732-3123(01)00051-7

diSessa, A. A. (2004). Metarepresentation: Native Competence and Targets for Instruction. *Cognition and Instruction, 22,* 293–331. http://dx.doi.org/10.1207/s1532690xci2203_2

Dreyfus, H., & Dreyfus, S. E. (1986). Five steps from novice to expert. In H. Dreyfus (Ed.), *Mind over machine: The power of human intuition and expertise in the era of the computer* (pp. 16–51). New York, NY: The Free Press.

Eilam, B., & Poyas, Y. (2008). Learning with multiple representations: Extending multimedia learning beyond the lab. *Learning and Instruction, 18,* 368–378. http://dx.doi.org/10.1016/j.learninstruc.2007.07.003

Gadgil, S., Nokes-Malach, T. J., & Chi, M. T. (2012). Effectiveness of holistic mental model confrontation in driving conceptual change. *Learning and Instruction, 22,* 47–61. http://dx.doi.org/10.1016/j.learninstruc.2011.06.002

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170. http://dx.doi.org/10.1207/s15516709cog0702_3

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52,* 45–56. http://dx.doi.org/10.1037/0003-066X.52.1.45

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York, NY: Prentice Hall.

Gibson, E. J. (2000). Perceptual learning in development: Some basic concepts. *Ecological Psychology, 12,* 295–302. http://dx.doi.org/10.1207/S15326969ECO1204_04

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition, 65,* 231–262. http://dx.doi.org/10.1016/S0010-0277(97)00047-4

Gutwill, J. P., Frederiksen, J. R., & White, B. Y. (1999). Making their own connections: Students' understanding of multiple models in basic electricity. *Cognition and Instruction, 17,* 249–282. http://dx.doi.org/10.1207/S1532690XCI1703_2

Hattie, J. (2012). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge.

Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior, 26,* 1246–1252. http://dx.doi.org/10.1016/j.chb.2010.03.025

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews, 6,* 53–84. http://dx.doi.org/10.1016/j.plrev.2008.12.001

Kellman, P. J., & Massey, C. M. (2013). Perceptual Learning, cognition, and expertise. *Psychology of Learning and Motivation, 58,* 117–165. http://dx.doi.org/10.1016/B978-0-12-407237-4.00004-9

Kellman, P. J., Massey, C. M., & Son, J. Y. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science, 2,* 285–305. http://dx.doi.org/10.1111/j.1756-8765.2009.01053.x

Kieren, T. E. (1993). Rational and fractional numbers: From quotient fields to recursive understanding. In T. P. Carpenter, E. Fennema, & T. A. Romberg (Eds.), *Rational numbers: An integration of research* (pp. 49–84). Hillsdale, NJ: Erlbaum.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science Education, 342,* 935–937. http://dx.doi.org/10.1126/science.1238056

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the Learning Sciences* (1 ed., pp. 61–77). New York, NY: Cambridge University Press.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36,* 757–798. http://dx.doi.org/10.1111/j.1551-6709.2012.01245.x

Kozma, R., & Russell, J. (2005). Students becoming chemists: Developing representationl competence. In J. Gilbert (Ed.), *Visualization in science education* (pp. 121–145). Dordrecht, Netherlands: Springer. http://dx.doi.org/10.1007/1-4020-3613-2_8

Lamon, S. J. (Ed.). (1999). *Teaching fractions and ratios for understanding.* Mahwah, NJ: Erlbaum.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science: A Multidisciplinary Journal, 11,* 65–100.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *Institute of Education Sciences, National Center for Special Education Research.* Retrieved from http://ies.ed.gov/ncser/pubs/20133000/on20133012/20133018/20132012

Martinie, S. L., & Bay-Williams, J. M. (2003). Investigating students' conceptual understanding of decimal fractions using multiple representations. *Mathematics Teaching in the Middle School, 8,* 244–247.

Moss, J., & Case, R. (1999). Developing children's understanding of the rational numbers: A new model and an experimental curriculum. *Journal for Research in Mathematics Education, 30,* 122–147. http://dx.doi.org/10.2307/749607

Moyer, P., Bolyard, J., & Spikell, M. A. (2002). What are virtual manipulatives? *Teaching Children Mathematics, 8,* 372–377.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence.* Reston, VA: National Council of Teachers of Mathematics.

National Mathematics Advisory Panel. (2008). *Foundations for success: Report of the National Mathematics Advisory Board Panel.* Washington, DC: U.S. Government Printing Office.

Ni, Y., & Zhou, Y.-D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist, 40,* 27–52. http://dx.doi.org/10.1207/s15326985ep4001_3

Ohlsson, S. (1988). Mathematical meaning and applicational meaning in the semantics of fractions and related concepts. In J. Hiebert & M. Behr (Eds.), *Research agenda for mathematics education: Number concepts and operations in the middle grades* (pp. 53–92). Reston, VA: National Council of Teachers of Mathematics.

Özgün-Koca, S. A. (2008). Ninth grade students studying the movement of fish to learn about linear relationships: The use of video-based analysis software in mathematics classrooms. *The Mathematics Educator, 18,* 15–25.

Pape, S. J., & Tchoshanov, M. A. (2001). The role of representation(s) in developing mathematical understanding. *Theory Into Practice, 40,* 118–127. http://dx.doi.org/10.1207/s15430421tip4002_6

Patel, Y., & Dexter, S. (2014). Using multiple representations to build conceptual understanding in science and mathematics. In M. Searson & M. Ochoa (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2014* (pp. 1304–1309). Chesapeake, VA: AACE.

Rau, M. A. (2016). Conditions for the effectiveness of multiple visual representations in enhancing STEM learning. *Educational Psychology*

Review. Advance online publication. http://dx.doi.org/10.1007/s10648-016-9365-3

Rau, M. A., Aleven, V., & Rummel, N. (2015). Successful learning with multiple graphical representations and self-explanation prompts. *Journal of Educational Psychology, 107,* 30–46. http://dx.doi.org/10.1037/a0037211

Rau, M. A., Aleven, V., Rummel, N., & Pardos, Z. (2014). How should Intelligent Tutoring Systems sequence multiple graphical representations of fractions? A multi-methods study. *International Journal of Artificial Intelligence in Education, 24,* 125–161. http://dx.doi.org/10.1007/s40593-013-0011-7

Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense making alone doesn't do it: Fluency matters too! ITS support for robust learning with multiple representations. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol. 7315, pp. 174–184). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-30950-2_23

Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2013). Why interactive learning environments can have it all: Resolving design conflicts between conflicting goals *Proceedings of the SIGCHI 2013 ACM Conference on Human Factors in Computing Systems* (pp. 109–118). New York, NY: ACM.

Raudenbush, S. W., & Bryk, A. S. (Eds.). (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Reimer, K., & Moyer, P. S. (2005). Third-graders learn about fractions using virtual manipulatives: A classroom study. *Journal of Computers in Mathematics and Science Teaching, 24*(1), 5–25.

Richman, H. B., Gobet, F., Staszewski, J. J., & Simon, H. A. (1996). Perceptual and memory processes in the acquisition of expert performance: The EPAM Model. In K. A. Ericsson (Ed.), *The road to excellence? The acquisition of expert performance in the arts and sciences, sports and games* (pp. 167–187). Mahwah, NJ: Erlbaum Associates.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49–70). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816819.005

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13,* 141–156. http://dx.doi.org/10.1016/S0959-4752(02)00017-8

Schooler, J. W., Fiore, S., & Brandimonte, M. A. (1997). At a loss from words: Verbal overshadowing of perceptual memories. *Psychology of Learning and Motivation: Advances in Research and Theory, 37,* 291–340. http://dx.doi.org/10.1016/S0079-7421(08)60505-8

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General, 122,* 166–183. http://dx.doi.org/10.1037/0096-3445.122.2.166

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction, 13,* 227–237. http://dx.doi.org/10.1016/S0959-4752(02)00022-1

Seufert, T., & Brünken, R. (2006). Cognitive load and the format of instructional aids for coherence formation. *Applied Cognitive Psychology, 20,* 321–331. http://dx.doi.org/10.1002/acp.1248

Shanks, D. (2005). Implicit learning. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 203–220). London: Sage. http://dx.doi.org/10.4135/9781848608177.n8

Siegler, R. S., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., & Wray, J. (2010). *Developing effective fractions instruction: A practice guide.* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Siegler, R. S., Fazio, L. K., Bailey, D. H., & Zhou, X. (2013). Fractions: The new frontier for theories of numerical development. *Trends in Cognitive Sciences, 17,* 13–19. http://dx.doi.org/10.1016/j.tics.2012.11.004

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62,* 273–296. http://dx.doi.org/10.1016/j.cogpsych.2011.03.001

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd. ed.). Cambridge, MA: MIT Press.

Stern, E., Aprea, C., & Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction, 13,* 191–203. http://dx.doi.org/10.1016/S0959-4752(02)00020-8

Taber, S. B. (2001). Making connections among different representations: The case of multiplication of fractions. Retrieved from http://eric.ed.gov/?id=ED454053

Thompson, D. R., & Saldanha, L. A. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick, G. Martin, & D. Schrifter (Eds.), *Research companion to the principles and standards for school mathematics* (pp. 95–114). Reston, VA: National Council of Teachers in Mathematics.

van der Meij, J., & de Jong, T. (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. *Learning and Instruction, 16,* 199–212. http://dx.doi.org/10.1016/j.learninstruc.2006.03.007

van der Meij, J., & de Jong, T. (2011). The effects of directive self-explanation prompts to support active processing of multiple representations in a simulation-based learning environment. *Journal of Computer Assisted Learning, 27,* 411–423. http://dx.doi.org/10.1111/j.1365-2729.2011.00411.x

Van Labeke, N., & Ainsworth, S. (2002). Representational decisions when learning population dynamics with an instructional simulation. In S. A. Cerri, G. Gouardères, & F. Paraguacu (Eds.), Proceedings of the 6th International Conference Intelligent Tutoring Systems (pp. 831–840). Berlin, Germany: Springer Verlag.

van Someren, M. W., Boshuizen, H. P. A., & de Jong, T. (1998). Multiple representations in human reasoning. In M. W. Van Someren, H. P. A. Boshuizen, & T. de Jong (Eds.), *Learning with multiple representations* (pp. 1–9). Oxford, United Kingdom: Pergamon Press.

Vreman-de Olde, C., & De Jong, T. (2006). Scaffolding learners in designing investigation assignments for a computer simulation. *Journal of Computer Assisted Learning, 22,* 63–73. http://dx.doi.org/10.1111/j.1365-2729.2006.00160.x

Wise, J. A., Kubose, T., Chang, N., Russell, A., & Kellman, P. J. (2000). Perceptual learning modules in mathematics and science instruction. In P. Hoffman & D. Lemke (Eds.), *Teaching and learning in a network world* (pp. 169–176). Amsterdam, the Netherlands: IOS Press.

Wylie, R., Koedinger, K. R., & Mitamura, T. (2009). Is self-explanation always better? The effects of adding self-explanation prompts to an English grammar tutor. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1300–1305). Amsterdam, the Netherlands.

Received August 13, 2015
Revision received May 25, 2016
Accepted June 3, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 3, 374–386

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000148

Conceptual Knowledge of Decimal Arithmetic

Hugues Lortie-Forgues
University of York

Robert S. Siegler
Carnegie Mellon University and Siegler Center for Innovative Learning, Beijing Normal University

In 2 studies (*Ns* = 55 and 54), the authors examined a basic form of conceptual understanding of rational number arithmetic, the direction of effect of decimal arithmetic operations, at a level of detail useful for informing instruction. Middle school students were presented tasks examining knowledge of the direction of effects (e.g., "True or false: $0.77 * 0.63 > 0.77$"), knowledge of decimal magnitudes, and knowledge of decimal arithmetic procedures. Their confidence in their direction of effect judgments was also assessed. The authors found (a) most students incorrectly predicted the direction of effect of multiplication and division with decimals below 1; (b) this pattern held for students who accurately compared the magnitudes of individual decimals and correctly executed decimal arithmetic operations; (c) explanations of direction of effect judgments that cited both the arithmetic operation and the numbers' magnitudes were strongly associated with accurate judgments; and (d) judgments were more accurate when multiplication problems involved a whole number and a decimal below 1 than with 2 decimals below 1. Implications of the findings for instruction are discussed.

Keywords: rational number arithmetic, decimal, conceptual knowledge, mathematical development, mathematical cognition

Supplemental materials: http://dx.doi.org/10.1037/edu0000148.supp

Understanding rational number arithmetic is central to a broad range of mathematical and scientific areas: algebra, geometry, trigonometry, statistics, physics, chemistry, biology, economics, and psychology, among them. One sign of this importance is that rational number arithmetic was part of more than half of the equations on the reference sheets for the most recent U.S. advanced placement physics and chemistry exams (College Board, 2014, 2015). Converging evidence comes from a longitudinal study of children's mathematics learning: In both the United Kingdom and the United States, 5th graders' fraction and decimal arithmetic performance predicted their algebra knowledge and overall mathematics achievement in tenth grade, even after IQ, socioeconomic status, race, ethnicity, whole number knowledge, reading comprehension, working memory, and other relevant variables were statistically controlled (Siegler et al., 2012). Beyond the classroom, rational number arithmetic is crucial for success not only in STEM areas but also in many occupations that do not require advanced math, including nursing, carpentry, and auto mechanic positions (e.g., Hoyles, Noss, & Pozzi, 2001; Sformo, 2008). This importance of rational number arithmetic both inside and outside the classroom is one reason why the Common Core State Standards Initiative (CCSSI, 2015) recommended that a substantial part of math instruction in 3rd through 7th grades be devoted to the subject.

Despite years of classroom instruction, many students fail to master arithmetic with decimals and fractions (Bailey, Hoard, Nugent, & Geary, 2012; Booth, Newton, & Twiss-Garrity, 2014; Byrnes & Wasik, 1991; Hecht, 1998; Hecht & Vagi, 2010; Hiebert & Wearne, 1985; Mazzocco & Devlin, 2008; Siegler, Thompson, & Schneider, 2011). Consider a few representative examples: (a) U.S. 8th graders who were tested on the four basic fraction arithmetic operations correctly answered only 57% of problems (Siegler & Pyke, 2013). (b) In a study of U.S. 9th graders, only 66% correctly answered the problem $4 + 0.3$, only 65% correctly answered $0.05 * 0.4$, and only 46% correctly answered $3 \div 0.6$ (Hiebert & Wearne, 1985). (c) On a standardized test with a nationally representative sample (the NAEP: National Assessment of Educational Progress) presented in 1978 and in a controlled experiment with the same item in 2014, fewer than 27% of U.S. 8th graders estimated correctly whether the closest answer to $12/13 + 7/8$ was 1, 2, 19, or 21 (Carpenter, Corbitt, Kepner, Lindquist, & Reys, 1980; Lortie-Forgues, Tian, & Siegler, 2015). (d) On the same NAEP, only 28% of U.S. 8th graders correctly chose whether the closest product to $3.04 * 5.3$, was 1.6, 16, 160, or 1,600 (Carpenter, Lindquist, Matthews, & Silver, 1983).

This article was published Online First August 15, 2016.

Hugues Lortie-Forgues, Department of Education, University of York; Robert S. Siegler, Department of Psychology, Carnegie Mellon University, and Siegler Center for Innovative Learning, Beijing Normal University.

The research reported here was supported in part by the Institute of Education Sciences, U. S. Department of Education, through Grants R305A150262 and R324C100004:84.324C, Subaward 23149 to Carnegie Mellon University, in addition to the Teresa Heinz Chair at Carnegie Mellon University, the Siegler Center of Innovative Learning, Beijing Normal University and a fellowship from the Fonds de Recherche du Québec – Nature et Technologies to Hugues Lortie-Forgues. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the Fonds de Recherche du Québec – Nature et Technologies.

Correspondence concerning this article should be addressed to Hugues Lortie-Forgues, Department of Education, Derwent College, University of York, York, YO10 5DD, United Kingdom. E-mail: hugues.lortie-forgues@york.ac.uk

The particular erroneous strategies that are used to solve rational number arithmetic problems convey the nature of the problem. With decimals, children often overgeneralize to multiplication the addition rule for placing the decimal point. They correctly answer that 1.23 + 4.56 = 5.79, but incorrectly claim that 1.23 * 4.56 = 560.88 (Hiebert & Wearne, 1985). Elementary, middle, and high school students also encounter difficulties when decimals involve one or more "0's" immediately to the right of the decimal point; many ignore those 0's and claim, for example, that 0.02 * 0.03 = 0.6 (Hiebert & Wearne, 1986). Similar erroneous strategies often appear with common fractions (i.e., numbers expressed as N/M), for example treating numerators and denominators as independent whole numbers and operating on them separately (e.g., 1/2 + 1/2 = 2/4; Ni & Zhou, 2005).

These and related data have led numerous investigators to suggest that students lack conceptual understanding of rational number arithmetic. Within this view, which we share, lack of understanding of rational number arithmetic limits students' ability to learn and remember the relevant procedures. For example, such lack of understanding could prevent students from rejecting implausible answers and the procedures that generated the answers and therefore lead the students not to search for more reasonable procedures.

Although the general conclusion is widely accepted, the specifics of what students do and do not understand about rational number arithmetic are largely unknown. Without these specifics, claims that students lack conceptual understanding have limited scientific use and few instructional implications. Therefore, the main purpose of the present study is to determine what middle school students do and do not understand about rational number arithmetic procedures, with an eye toward specifying the difficulties at a level useful for improving instruction.

In Study 1, we examined whether a particularly striking type of misunderstanding—direction of effect errors—are seen with decimals, as they previously have been documented to be with common fractions. We also examined children's confidence ratings of their direction of effect judgments to distinguish among several theoretical interpretations of the judgments. In Study 2, we determined whether direction of effect misconceptions extend to problems involving a whole number and a decimal and also obtained explanations of direction of effect judgments to better understand the reasoning underlying children's judgments.

Direction of Effect of Rational Number Arithmetic Operations

Perhaps the most basic understanding about rational number arithmetic is the direction of effect that the operations produce: Will the answer be larger or smaller than the operands (the numbers in the problem). To examine knowledge of this type, Siegler and Lortie-Forgues (2015) devised a direction of effect task that presented inequalities such as the following: "True or False: 31/56 * 17/42 > 31/56." Fractions with relatively large numerators and denominators were used to prevent participants from calculating exact answers and thus answering correctly without considering the direction of effect of the arithmetic operation with those numbers.

For addition and subtraction of positive numbers, the direction is the same regardless of the size of the numbers: addition of positive numbers always yields an answer greater than either operand, and subtraction always yield an answer smaller than the number from which another number is being subtracted. However, for multiplication and division, the direction of effect varies with the size of the operands. Multiplying numbers above 1 always yields a product greater than either multiplicand, but multiplying numbers between 0 and 1 never does. Conversely, dividing by numbers above 1 always results in answers less than the number being divided, but dividing by numbers between 0 and 1 never does. Without understanding these relations, people cannot evaluate an answer's plausibility.

The implausible answers to rational number arithmetic problems that many students generate might be taken as evidence that students lack direction of effect knowledge. However, such answers might reflect students focusing on executing the computations and not considering the answer's plausibility. Rational number arithmetic imposes a high working memory load (English & Halford, 1995), which could lead to students not considering answers' plausibility. Therefore, to examine whether people reveal understanding of the direction of effect of fraction arithmetic when freed from the processing load imposed by computing, Siegler and Lortie-Forgues (2015) presented addition, subtraction, multiplication, and division direction of effect problems with fractions above 1 and fractions below 1 to 6th and 8th graders (12- and 14-year-olds) and preservice teachers attending a highly ranked school of education.

The most striking finding of the study was that 6th graders, 8th graders, and preservice teachers all were below chance in judging the direction of effect of multiplying and dividing fractions below 1. For example, preservice teachers erred on 67% of trials, and middle school students on 69% when asked to predict whether multiplying two fractions below 1 would produce an answer larger than the larger operand. These findings did not reflect weak knowledge of the fraction arithmetic procedures. The pattern was present even among the many preservice teachers and children whose fraction arithmetic computation was perfect for the same operation, indicating that the inaccurate direction of effect judgments were not attributable to the teachers and students not knowing the relevant arithmetic operations. This observation attests to people being able to memorize mathematical procedures without even the most basic understanding of them. The findings also did not mean that the task was confusing or impossible. Math and science majors at a selective university erred on only 2% of the same problems.

These findings were not idiosyncratic to the task or samples. Highly similar findings emerged on a related item from the 2011 TIMSS (Trends in International Mathematics and Science Study), a standardized international comparison of math knowledge (Mullis, Martin, Foy, & Arora, 2012). Eighth graders were asked to judge which of four locations on a number line included the product of two unspecified fractions below 1. The locations were (a) between 0 and the smaller multiplicand, (b) between the two multiplicands, (c) between the larger multiplicand and 1, and (d) halfway between 1 and 2. Consistent with Siegler and Lortie-Forgues' (2015) findings, 77% of U.S. 8th graders erred on the problem.

These findings from both the experimental study and the large-sample international assessment raise the issue of whether difficulties understanding direction of effect of rational number arith-

metic procedures are limited to fraction arithmetic or whether they reflect a more general difficulty in understanding multiplication and division of rational numbers, one that extends to decimals as well as fractions. It was entirely plausible that the difficulty with direction of effect judgments was limited to fractions. Fraction notation seems likely to (a) make it difficult to accurately estimate the magnitudes of individual numbers, which (b) increases the difficulty of estimating answers to arithmetic problems using those numbers, which (c) makes it difficult to recognize the implausibility of many answers yielded by incorrect fraction arithmetic procedures, which (d) makes it difficult to rule out these incorrect procedures, thereby reducing searches for correct procedures.

Consistent with the idea that fraction notation makes estimation of individual number's magnitude difficult, 8th graders' estimates for fractions between 0 and 5 are less accurate than second graders' estimates for whole numbers between 0 and 100 (Laski & Siegler, 2007; Siegler et al., 2011). The greater difficulty of accurately estimating fraction magnitudes is unsurprising, because a fraction's magnitude must be derived from the ratio of the numerator and denominator rather than from a single number, as with whole numbers and decimals. Consistent with the idea that the fraction notation increases the difficulty of estimating answers to fraction arithmetic problems, middle school students are very inaccurate in estimating the answers to fraction arithmetic problems (Hecht & Vagi, 2010). Finally, consistent with the ideas that fraction notation makes it difficult to recognize implausible answer and rule out the wrong procedures that generated them, children frequently generate implausible fraction arithmetic answers, both through treating numerators and denominators as independent whole numbers ($1/2 + 1/2 = 2/4$) and through only operating on the numerator ($12/13 + 7/8 \cong 19$) (Lortie-Forgues et al., 2015; Ni & Zhou, 2005). Thus, inaccuracy on the direction of effect task with fraction multiplication and division in Siegler and Lortie-Forgues (2015) and on the related TIMSS item might have reflected difficulties specific to fractions, especially difficulty accessing fraction magnitudes.

Another possibility, however, is that the inaccurate direction of effect judgments with fraction multiplication and division might reflect poor understanding of multiplication and division that extends beyond fractions and that has nothing directly to do with lack of magnitude understanding of individual numbers. In particular, participants might have overgeneralized the pattern of answers from whole number arithmetic and not understood that there is nothing about multiplication that requires answers to be greater than either operand and nothing about division that requires answers to be less than the number being divided. This interpretation suggests that weak understanding of multiplication and division should be as evident with decimals as with the corresponding fractions.

Supporting this latter interpretation, overgeneralizations from whole to rational numbers are very common with decimals, common fractions, and negatives alike. When comparing the magnitude of individual decimals, children often think that, as with whole numbers, more numerals implies larger magnitudes (e.g., claiming that $.35 > .9$; Resnick et al., 1989; Resnick & Omanson, 1987). Similarly, many children err on fraction magnitude comparison problems by assuming that fractions with larger whole number values for numerators and denominators are larger than fractions with smaller ones (e.g., $11/21 > 3/5$; Fazio, Bailey, Thompson & Siegler, 2014; Ni & Zhou, 2005). Overgeneralization of whole number knowledge is also common with negative numbers (e.g., $-12 > -6$; Ojose, 2015).

Examining direction of effect judgments for decimal arithmetic provided a means for contrasting these two explanations. Unlike fractions, decimals are expressed by a single number, a feature that facilitates access to decimal magnitudes. To appreciate the difference, contrast the difficulty of judging the relative sizes of 7/9 and 10/13 with the ease of judging the relative sizes of their decimal equivalents, 0.78 and 0.77. Empirical data support this analysis; magnitude comparisons of college students are much faster and more accurate with decimals than fractions (DeWolf, Grounds, Bassok, & Holyoak, 2014). The same pattern holds for number line estimation as for magnitude comparison, and for children as well as adults (Iuculano & Butterworth, 2011; Desmet, Gregoire, & Mussolin, 2010).

Thus, if the inaccurate direction of effect judgments with multiplication and division of fractions between 0 and 1 was due to difficulty accessing fraction magnitudes, then presenting the same task with decimals should reduce or eliminate the difficulty. If magnitude knowledge influenced direction of effect judgments, we also would expect individual children's accuracy on measures of the two types of knowledge to correlate positively. On the other hand, if inaccurate direction of effect judgments reflected limited understanding of multiplication and division, the same pattern should be evident with decimals as with fractions.

Our prediction was that the same difficulties with judging direction of effect for multiplication and division of operands between 0 and 1 would be present with decimals as had been documented previously with fractions. One source of support for this prediction was that when 4th and 5th graders were asked for their reaction to being told that $15 * 0.6 = 9$, many children expressed surprise, with 25% saying without prompting that they expected the answer to be larger than 9 (Graeber & Tirosh, 1990). Similar reactions were observed in the same study when students were told that $12 \div 0.6 = 20$. Another paradigm has yielded similar results: When presented operands and answers and asked to select the appropriate operation, both high school students and preservice teachers generally chose multiplication when problems yielded answers larger than the numbers being multiplied, and they chose division when problems yielded answers smaller than the number being divided, regardless of the semantics of the problem (Fischbein, Deri, Nello, & Marino, 1985; Tirosh & Graeber, 1989). Moreover, in previous studies of decimal arithmetic, students have been found to often misplace the decimal point on multiplication and division problems in ways that reflected little understanding of the plausibility of the answer (Hiebert & Wearne, 1985, 1986).

However, there was reason to hope that these findings underestimated current students' conceptual understanding of decimal arithmetic. One consideration was that the prior findings with decimals are 25 or more years old; the increased educational emphasis on conceptual understanding of rational numbers in recent years (e.g., CCSSI, 2015) might have increased understanding of decimal arithmetic among contemporary students. Moreover, the prior findings might underestimate children's understanding of decimal arithmetic: the participants tested had either had very little experience with decimal arithmetic (Graeber & Tirosh, 1990) or the questions consisted of

word problems, which often require complex verbal processing in addition to mathematical understanding (Fischbein et al., 1985; Tirosh & Graeber, 1989).

A second purpose of Study 1 was to examine students' confidence in their direction of effect judgments. On mathematics problems, people sometimes generate wrong answers that they believe are correct; at other times, they generate wrong answers that they doubt are correct but cannot generate more likely alternatives. Participants in Siegler and Lortie-Forgues (2015) might have been convinced that their incorrect direction of effect judgments were correct, but they might have been unsure and relied on their whole number knowledge as a default option because they did not know what else to do. This type of default explanation seems to be common when people have limited knowledge of a topic (see Rozenblit & Keil, 2002 for examples of default explanations in nonmathematical contexts).

Obtaining confidence ratings allowed us to distinguish among three theoretical interpretations of incorrect direction of effect judgments on multiplication and division with decimals below 1: (a) The *strong conviction hypothesis*, which posits that students are highly confident that multiplication produces answers greater than either operand and division produces answers less than the dividend; (b) The *operation knowledge hypothesis*, according to which students recognize that they know less about multiplication and division than addition and subtraction, and therefore are less confident in their multiplication and division judgments, regardless of whether the operands are below or above 1; (c) The *cognitive conflict hypothesis*, in which, due to the contradiction between children's whole number experience and their experience multiplying and dividing numbers between 0 and 1, they are less confident in their multiplication and division direction of effect judgments with numbers between 0 and 1 than in their other judgments.

If the strong conviction hypothesis is correct, confidence ratings for all eight types of problems should be equally high. If the operation knowledge hypothesis is correct, confidence ratings for the four addition and subtraction problems should be higher than for the four multiplication and division problems. If the cognitive conflict interpretation is correct, confidence ratings for multiplication and division of decimals below 1 should be lower than for the other six types of problems. Combinations of these alternatives were also possible; for example, children might be less confident in their multiplication and division judgments on all problems, and especially unconfident of judgments when those operations involve operands between 0 and 1.

Study 1

Method

Participants. The children were 55 middle school students (19 sixth and 36 eighth graders; 27 boys, 28 girls, *M* age = 12.75 years, *SD* = 1.06) who attended a public school in a middle-income suburban area near Pittsburgh, PA. These age groups were chosen because decimals were taught in the children's schools in 5th and 6th grades prior to the study and because doing so allowed direct comparison between direction of effect knowledge for fractions, which was examined in Siegler and Lortie-Forgues (2015), and for decimals, which was examined here. The school district included 59% Caucasian, 35% African American, 1% Asian, and 5% "other" children. Math achievement test scores were average for the state; 76% of 6th graders and 81% of 8th graders were at or above grade level, versus 78% and 75% for the state. Students were tested in groups in their math classroom during a regular class period in the middle of their school year.

Tasks.

Direction of effect judgments and confidence ratings. This task included 16 mathematical inequalities, four for each arithmetic operation. Each item was of the form "True or false: $a * b > a$?" Both a and b were two-place decimals, and a was always larger than b. On half of the problems, both a and b were below 1 (e.g., $0.77 * 0.63 > 0.77$); on the other half, both were above 1 (e.g., $1.36 * 1.07 > 1.36$). The same pairs of operands—0.77 and 0.63, 0.94 and 0.81, 1.36 and 1.07, and 1.42 and 1.15—were presented with all four arithmetic operations. Four problems, one with each arithmetic operation, were presented on each page of a booklet that children received; each pair of operands was used once on each page. Students received one point for each correct judgment.

After each problem, children were asked to rate their confidence in their answer on a 5-point scale ranging from 1 (*not confident at all*) to 5 (*extremely confident*). The numerical value of each confidence rating constituted the data on that trial; effects of arithmetic operation and operand size (above or below 1) on the confidence ratings were analyzed.

Arithmetic computation. Participants were asked to answer 12 computation problems, 3 for each arithmetic operation. For each arithmetic operation, the operand pairs were 0.9 and 0.4, 0.45 and 0.18, and 3.3 and 1.2. The task was included to examine whether computation skill was related to understanding of direction of effects of the arithmetic operations.

Magnitude comparison. Children were presented 32 problems requiring comparison of 0.533 to another decimal. Half of the decimals were larger and half smaller than 0.533; equal numbers of these comparison numbers had 1, 2, 3, or 4 digits to the right of the decimal.

Procedure. Tasks were always presented in the following order: direction of effect, arithmetic computation, and magnitude comparison. Items within each task were presented in one of two orders, either first to last or last to first. All tasks were presented in printed booklets, with students writing answers with pencils. Students were asked to perform the problems in order; use of calculators was not allowed. The experiment was conducted by two research assistants and Hugues Lortie-Forgues.

Reliabilities. Reliabilities of the measures (Cronbach's alpha) were above the satisfactory value of 0.70 (Nunnally, 1978), except in cases where ceiling effects were present, a factor known to lower reliabilities (May, Perez-Johnson, Haimson, Sattar, & Gleason, 2009). One case where ceiling effects were present and appeared to lower reliability involved the internal consistency of direction of effect judgments. The relatively low coefficient alpha on this task, $\alpha = .68$, appeared to be due to a ceiling effect on problems where performance was highly accurate and therefore where there was little variability. These were problems involving all four arithmetic operations when operands were above 1 and addition and subtraction problems with operands below 1. More than half of students (56%) were 100% accurate on these 12 problems. On direction of effect problems where performance

varied to a greater extent (multiplication and division of numbers below 1), internal consistency was adequate (α = .74 and 0.80, respectively). Low internal consistency on the arithmetic computation task, α = .58, also appeared due to ceiling effects. In this case 64% of students correctly answered all addition and subtraction computation problems. Again, internal consistency on multiplication and division computation problems, where performance was more variable, was adequate (α = .71 and 0.76, respectively). Reliability of confidence ratings for direction of effect judgments was high (α = .95), as was internal consistency of magnitude comparisons (α = .94). See online supplemental Table S1 for the results presented separately for each grade on each task.

Results and Discussion

Direction of effect judgments. A repeated-measures analysis of variance (ANOVA) with decimal size (above or below 1) and arithmetic operation (addition, subtraction, multiplication, or division) as within-subject factors, grade (6th or 8th) as a between-subjects factor, and number of correct direction of effect judgments as the dependent variable yielded main effects of arithmetic operation, $F(3, 159) = 52.61$, $p < .001$, $\eta p^2 = 0.49$, and decimal size, $F(1, 53) = 63.02$, $p < .001$, $\eta_p^2 = 0.54$, as well as a Decimal Size × Arithmetic Operation interaction, $F(3, 159) = 38.24$, $p < .001$, $\eta_p^2 = 0.42$. Post hoc comparisons with the Bonferroni correction showed that number of correct predictions for decimals below and above 1 did not differ on addition (87% vs. 88% correct; $t[54] = 0.63$, $p = .53$) or subtraction (89% vs. 90%; $t[54] = 0.29$, $p = .77$), but differed greatly on multiplication (20% vs. 84%; $t[54] = 7.12$, $p < .001$) and division (19% vs. 89%; $t[54] = 9.04$, $p < .001$). Accuracy was below the chance level of 50% with decimals below 1 for both multiplication, $t[54] = 6.05$, $p < .001$ and division, $t[54] = 6.49$, $p < .001$.

Analysis of individual children's judgments showed similar findings. Half (49%) of students erred on all four multiplication and division problems with operands below 1 and correctly answered all 12 other problems.

As shown in Table 1, these direction of effect judgments with decimals mirrored previous data with fractions, with the single exception that decimal division judgments with operands below 1 were *less* accurate than the corresponding fraction judgments. The parallel patterns suggest that students' performance reflected a misunderstanding of multiplication and division that is independent of the numbers' format (see online supplemental Table S2 for the percentages for each grade reported separately).

Table 1
Percent Correct Direction of Effect Judgments for Decimal and Fraction Arithmetic by Operand Size and Arithmetic Operation

Operand sizes	Operation	Decimals	Fractions
Above one	Addition	88	92
	Subtraction	90	94
	Multiplication	84	92
	Division	89	70
Below one	Addition	87	89
	Subtraction	89	92
	Multiplication	20	31
	Division	19	47

Note. Percentages for fraction arithmetic in the right hand column are from grade peers in Siegler & Lortie-Forgues, 2015.

Confidence ratings. Confidence ratings for the direction of effect task were analyzed via a parallel repeated-measures ANOVA with decimal size and arithmetic operation as within-subject factors and grade as a between-subjects factor. The analysis yielded a main effect of arithmetic operation, $F(3, 159) = 20.31$, $p < .001$, $\eta_p^2 = 0.28$, and a Decimal Size × Grade interaction, $F(1, 53) = 4.63$, $p = .036$, $\eta_p^2 = 0.08$. Post hoc comparisons with the Bonferroni correction showed that confidence in direction of effect judgments was lower for division ($M = 3.97$, $SD = 1.01$) than for multiplication ($M = 4.37$, $SD = 0.72$; $t[54] = 4.24$, $p < .001$), and lower for multiplication than for addition ($M = 4.53$, $SD = 0.63$; $t[54] = 2.61$, $p = .01$) or subtraction ($M = 4.56$, $SD = 0.60$; $t[54] = 3.17$, $p < .01$). The Decimal Size × Grade interaction reflected 8th but not 6th graders being less confident in their judgments on problems with decimals below 1 than on problems with decimals above 1—for 8th graders, mean confidence rating of 4.31 vs. 4.41, $t(35) = 2.64$, $p = .01$; for 6th graders, mean rating of 4.38 vs. 4.34, $t(18) = 0.77$, $p = .45$.

We next examined confidence ratings of the half of participants (49%) whose judgments always matched the direction of effect of arithmetic with whole numbers (i.e., always wrong on the 2 multiplication and 2 division problems with decimal operands below 1 and always correct on the other 12 problems). The analysis yielded a main effect of arithmetic operation, $F(3, 75) = 11.48$, $p < .001$, $\eta_p^2 = 0.315$. Confidence in direction of effect judgments was lower for division ($M = 4.05$, $SD = 1.09$) than for addition ($M = 4.59$, $SD = 0.44$; $t[26] = 3.29$, $p = .002$), subtraction ($M = 4.58$, $SD = 0.48$; $t[26] = 3.60$, $p < .001$), and multiplication ($M = 4.58$, $SD = 0.53$; $t[26] = 3.46$, $p = .002$). Confidence ratings did not differ between problems with numbers above and below 1 (for problems above 1, $M = 4.44$, $SD = 0.58$; for problems below 1, $M = 4.46$, $SD = 0.58$; $t[26] = 0.42$, $p = .7$).

In contrast, conducting the same analysis on the 51% of participants whose judgments did not invariably follow the direction of effect of whole number arithmetic yielded a decimal size X grade interaction, $F(1, 26) = 7.92$, $p = .009$, $\eta_p^2 = 0.233$, as well as a main effect of arithmetic operation, $F(3, 78) = 11.25$, $p < .001$, $\eta_p^2 = 0.302$. The main effect reflected lower confidence in division judgments ($M = 3.89$, $SD = 0.95$) than in ones for multiplication ($M = 4.17$, $SD = 0.82$; $t[27] = 2.51$, $p = .018$), and for multiplication judgments than for addition (M = 4.46, $SD = 0.77$; $t[27] = 3.21$, $p = .003$) and subtraction ($M = 4.54$, $SD = 0.71$; $t[27] = 4.47$, $p < .001$) ones. The interaction arose from 8th graders being less confident in their judgments on problems with decimals below than above 1 (Ms = 4.09 and 4.27, SDs = .84 and .80; $t[18]$ 2.96, $p = 0.008$), but no difference being present for 6th graders (Ms = 4.50 vs. 4.38, $t[8]$ 1.37, $p = .206$). This interaction suggested that by 8th grade, children began to recognize that there was something different about computations with decimals below 1 than decimals above 1.

Arithmetic computation. A repeated-measures ANOVA on accuracy of decimal arithmetic computation, with arithmetic operation as a within-subject factor, grade as a between-subjects factor, and number of correct answers as the dependent variable yielded a main effect of arithmetic operation, $F(3, 159) = 51.5$, $p < .001$, $\eta_p^2 = 0.493$. Post hoc comparisons with the Bonferroni correction showed that number correct was lower on division

problems (M = 35%, SD = 40%) than on multiplication problems (M = 54%, SD = 39%, t[54] = 3.08 $p < .01$) and lower on multiplication than on addition (M = 90%, SD = 18%, t[54] = 5.98, $p < .001$) and subtraction problems (M = 93%, SD = 18%, t[54] = 6.23, $p < .001$). There was no effect of grade, but 8th graders tended to generate more correct answers on multiplication (6th graders 44%; 8th graders 59%) and division (6th graders 21%; 8th graders 43%) problems.

Decimal arithmetic accuracy (68% correct) closely resembled that on similar problems 30 years ago (e.g., Hiebert & Wearne, 1985). Also as then, misplacing the decimal point in the answer was the most common source of multiplication errors. On 73% of multiplication errors (34% of answers), students multiplied correctly but misplaced the decimal in the answer. Misplacing the decimal was also a fairly frequent source of division errors (21% of errors; 13% of answers).

The below chance direction of effect judgment accuracy on multiplication and division of decimals below 1 was not attributable to the less accurate computation on those operations. Most students (14 of 19, 74%) who correctly solved both multiplication computation problems involving decimals below 1 were incorrect on both of the direction of effect judgments on parallel problems. Similarly, among students who correctly answered both of the division computation problems with decimals below 1, most (nine of 14, 64%) erred on both of the corresponding direction of effect problems.

For both 6th and 8th graders, numbers of correct arithmetic computations and direction of effect judgments were weakly correlated or uncorrelated (6th grade, $r = -0.28$, *ns*; 8th grade, $r = .33$, $p = .05$). The pattern was similar when the problems of greatest interest were analyzed separately. No relation was present when only multiplication direction of effect problems with operands below 1 and multiplication computation problems with operands below 1 were considered (6th grade, $r = .26$, *ns*; 8th grade, $r = .13$, *ns*) or when only division direction of effect problems with operands below 1 and division computation problems with operands below 1 were considered (6th grade, $r = .37$, *ns*; 8th grade, $r = .19$, *ns*).

Magnitude comparison. Children correctly answered 83% of decimal magnitude comparisons. Performance was higher when the two decimals being compared had the same number of decimal places than when they had different numbers of decimal places (90% vs. 80% correct, t[54] 3.34, $p = .002$). Accuracy did not differ significantly between 6th and 8th graders, 77% versus 86%, $t(54) = 1.53$, $p = .13$.

Analyses of magnitude comparison errors showed large individual differences in knowledge of decimal magnitudes. At one extreme, 53% of children correctly answered more than 95% of decimal comparisons. At the other extreme, 18% of children answered incorrectly more than 90% of the 12 items on which ignoring the decimal point yielded a wrong answer (e.g., saying that 0.9 is smaller than 0.533, because $9 < 533$).

For both 6th and 8th graders, numbers of correct magnitude comparison and direction of effect judgments were unrelated (6th grade, $r = .01$, *ns*; 8th grade, $r = .10$, *ns*). The same was true when only multiplication direction of effect problems with operands below 1 were considered (6th grade, $r = .17$, *ns*; 8th grade, $r = -0.05$, *ns*) and when only division direction of effect problems with operands below 1 were (6th grade, $r = -0.03$, *ns*; 8th grade, $r = .10$, *ns*).

In summary, direction of effect judgments with decimals were much like those observed by Siegler and Lortie-Forgues (2015) with common fractions. The 6th and 8th graders erred more often than chance on problems involving multiplication and division of decimals below 1 but were highly accurate on all other types of problems. These results with decimals could not be attributed to lack of magnitude knowledge. With both problems in general and on the two types of problems that elicited inaccurate direction of effect judgments, accuracy of magnitude comparison performance and direction of effect judgments were at most weakly related.

Study 2 was designed to build on these findings by examining direction of effect judgments on a type of problem that was potentially important for instruction—problems that include a whole number and a decimal. Such problems provide a possible transition context through which instruction could build on students' understanding of whole number arithmetic and extend it to decimals. Study 2 also was designed to deepen our understanding of children's thinking about direction of effect judgments by having them explain their reasoning on them. As will be seen, the explanations proved invaluable for demonstrating that accurate predictions sometimes reflect processes quite different than the ones on which the predictions were based.

Study 2

In some U.S. textbooks series, such as *Everyday Math* (Bell et al., 2007) and *Prentice Hall Mathematics* (Charles, Illingworth, McNemar, Mills, & Ramirez, 2012), problems involving a whole number and a decimal below 1 are presented quite often. A likely reason is that such problems can capitalize on students' familiarity with whole numbers and with the usual framing of whole number multiplication as repeated addition. For instance, 5 * 0.34 can be interpreted as five iterations of 0.34. Even the phrasing "5 times 0.34" supports this interpretation. In contrast, the repeated addition interpretation is difficult to apply to multiplication if both operands are below 1 (viewing 0.05 * 0.3 as 0.3 being added 0.05 times is less intuitive than viewing 5 * 0.03 as 0.3 being added 5 times).

Because the repeated addition interpretation applies more straightforwardly to multiplication problems with a whole number and a decimal (WD problems) than to problems with two decimals (DD problems), direction of effect judgments for multiplication might be more accurate on WD than DD problems. Children could solve direction of effect problems with a whole and a decimal below 1 by estimating the result of adding the decimal the whole number of times; this logic is much more difficult to apply to problems with two decimals. However, students might not use the repeated addition interpretation of multiplication on either type of problem, because they were so convinced that multiplication always produces answers larger than the operands that they did not consider other possibilities, because they did not think of the repeated addition interpretation, or because they relied on some other interpretation. Thus, one goal of Experiment 2 was to test whether direction of effect judgments were more accurate on WD than DD multiplication problems.

At first glance, the same logic would seem to apply to division. For example, 3 ÷ 0.5 could be solved by six additions of 0.5, and children could solve the corresponding direction of effect problem

by estimating the number of times 0.5 would need to be added to reach 3. However, several considerations suggested that for division, direction of effect problems would be no easier on WD than on DD problems. Although repeated addition and subtraction can be used to solve some WD division problems (ones where the dividend is bigger than the divisor and that have a whole number answer), the most common interpretation of division appears to be equal sharing (Carpenter, Illingworth, McNemar, Mills, & Ramirez, 1999; Rizvi & Lawson, 2007). That interpretation makes sense with whole numbers (e.g., 30 ÷ 3 means 30 cookies shared equally among three friends) but is meaningless with decimal divisors (e.g., what does it mean to share 30 cookies among 0.3 friends). Because the equal sharing interpretation is not easily applicable to problems with decimal divisors, and because the repeated addition interpretation is useful for understanding only on a subset of division WD problems, we did not expect a difference between direction of effect judgment accuracy on WD and DD division problems.

A second main goal of Study 2 was to deepen our analysis of conceptual understanding of rational number arithmetic by asking students to explain the reasoning underlying their judgments on the direction of effect task. We were particularly interested in testing whether they apply the logic of repeated addition more often to WD than DD multiplication problems, and whether this logic underlay the predicted greater accuracy on WD than DD problems.

Method

Participants. Participants were 54 seventh graders (26 boys, 28 girls, *M* age = 12.7 years, *SD* = 0.54) who attended a public school in a middle-income suburban area near Pittsburgh, PA. The school district included 63% Caucasian, 22% African American, 7% Asian, 2% Hispanic, and 7% "other" children. As in Experiment 1, the school's mean math achievement was similar to that in the state as a whole (79% of 7th graders in the district were at or above grade level, 73% in the state). Students were tested in groups in their math classroom during a regular class period near the end of the school year. A research assistant and a postdoctoral student (Hugues Lortie-Forgues) collected the data.

Tasks.

Direction of effect judgment only task. Each student was presented 36 problems (18 DD and 18 WD items). For each type of problem, there were six addition, six multiplication and six division items. Subtraction items were not presented to reduce the duration of the experiment and because direction of effect judgments on addition and subtraction problems were almost identical in the previous experiment.

Half of the DD problems for each operation involved pairs of decimals below 1; the other half involved pairs of decimals above 1. All WD problems for each operation included a whole number above 1; half of these items included a decimal below 1 and half a decimal above 1. On all WD problems, the whole number appeared first, the decimal appeared second, and the comparison answer was the whole number (e.g., "True or false: 5 * 0.291 > 5"). Problems were generated using one of the following sets of operand pairs:

Set A DD problems: 0.87 and 0.291; 0.96 and 0.173; 0.79 and 0.356; 8.83 and 3.584; 6.14 and 5.781; 12.87 and 2.854;

Set A WD problems: 5 and 0.291; 4 and 0.173; 14 and 0.356; 8 and 3.584; 6 and 5.781; 12 and 2.854.

Set B DD problems: 0.76 and 0.182; 0.85 and 0.261; 0.97 and 0.345; 9.74 and 5.495; 7.26 and 3.853; 11.49 and 2.898;

Set B WD problems: 6 and 0.182; 8 and 0.261; 13 and 0.345; 9 and 5.495; 7 and 3.853; 11 and 2.898).

DD problems were presented consecutively, as were WD problems. Problem order (DD problems first or WD problems first) and problem set (DD problems from set A and WD problems from set B, or vice versa) were counterbalanced. The items in Set A and Set B were chosen to be as similar as possible.

Judgment plus explanation task. The format of this task was identical to that of the judgment only task, except that students were asked to explain their reasoning immediately after each judgment. Such immediately retrospective strategy reports have been found to yield valid and nonreactive data for many numerical tasks, including arithmetic and number line estimation (e.g., Siegler, 1987; Siegler et al., 2011). Presenting both the judgment only task and the judgment plus explanation task allowed us to obtain explanations data and also to test whether obtaining explanations affected judgments.

Each student was presented with 12 judgment plus-explanation problems (6 DD and 6 WD problems; two addition, two multiplication, and two division problems within each group; half with operands above 1, and half with operands below 1). Each problem was generated using one of two sets of operand pairs:

Set A DD items: 0.87 and 0.291; 8.83 and 3.584;

Set A WD items: 5 and 0.291; 8 and 3.584;

Set B DD Items: 0.76 and 0.182; 9.74 and 5.495;

Set B WD Items: 6 and 0.182; 9 and 5.495.

For each participant, order of problems (DD or WD first) was the same as on the judgment only task, but the sets of operand pairs used to generate the problems were switched. Participants whose DD problems on the judgment-only task were from Set A were presented DD problems on the judgment-plus-explanation task from Set B, and vice versa.

Magnitude comparison. The task was the same as in Experiment 1, except that the problems where the decimals being compared had the same number of decimal places were excluded. This resulted in 24 decimal magnitude comparison problems.

Procedure. The three tasks were presented in booklets in the following order: (a) direction of effect judgment-only task, (b) direction of effect judgment-plus-explanation task, (c) magnitude comparison task. Children were asked to complete the tasks without a calculator in the order in which they appeared in the booklet.

Reliabilities of measures. Measures of internal consistency (Cronbach's alpha) of the direction of effect judgment only task, the judgment plus explanation task, and the magnitude comparison task were all *satisfactory (α = .74, 0.71 and 0.95, respectively).*

Results and Discussion

Direction of effect judgment-only task. We computed a repeated-measures ANOVA with decimal size (above or below 1),

arithmetic operation (addition, multiplication, or division) and whole number operand (present or absent) as within-subject factors; problem set (A or B) and problem order (DD first or WD first) as between-subjects factors; and number of correct judgments as the dependent variable.

Main effects emerged for arithmetic operation, $F(2, 88) = 80.21$, $p < .001$, $\eta_p^2 = 0.646$, and decimal size, $F(1, 44) = 79.43$, $p < .001$, $\eta_p^2 = 0.644$. Three interactions also were present: Arithmetic Operation × Whole Number Operand, $F(2, 88) = 3.49$, $p = .035$, $\eta_p^2 = 0.073$; Arithmetic Operation × Decimal Size, $F(2, 88) = 39.80$, $p < .001$, $\eta_p^2 = 0.475$; and Arithmetic Operation × Whole Number Operand × Decimal Size, $F(2, 88) = 3.48$, $p = .035$, $\eta_p^2 = 0.073$.

The three-way interaction and the two two-way interactions could be interpreted quite straightforwardly. As shown in the three rows at the top of Table 2, when both operands were above 1, answers were uniformly accurate on all three arithmetic operations. Neither arithmetic operation nor presence of a whole number affected accuracy on these problems. The high accuracy seems attributable to the direction of effect being the same for decimals as for whole numbers.

As shown in the three rows at the bottom of Table 2, the pattern differed with decimals below 1. On these problems, addition judgments were accurate and division problems inaccurate regardless of whether the problem included a whole number. These findings also appeared due to generalization from effects of the operation with whole numbers. In contrast, and consistent with our prediction, on multiplication problems with decimals below 1, direction of effect judgments were more accurate when one operand was a whole number (WD problems) ($M = 47\%$, $SD = 47\%$) than when both operand were decimals (DD problems) ($M = 31\%$, $SD = 42\%$), $t(53) = 2.97$, $p < .01$, Cohen's $d = 0.41$. This pattern was consistent across problems; direction of effect judgments were more accurate on all three multiplication problems that involved a whole number and a decimal below 1 (43%-50% correct) than on any of the three multiplication problems that involved two decimals below 1 (30–33% correct). Consistent with this interpretation, accuracy with decimals below 1 was below the chance level (i.e., 50%) on multiplication DD problems, $t(53) = 3.35$, $p < .001$; division DD problems, $t(53) = 6.19$, $p < .001$; and division WD problems, $t(53) = 5.52$, $p < .001$; but not on multiplication WD problems, $t(53) = 0.48$, $p = .63$.

Table 2
Percent Correct Judgments on the Direction of Effect Judgments Task and on the Judgments Plus Explanations Task

Operand size	Operation	Judgments task		Judgments plus explanations task	
		DD problems	WD problems	DD problems	WD problems
Above one	Addition	97	96	94	96
	Multiplication	92	92	98	96
	Division	94	95	96	100
Below one	Addition	96	94	94	91
	Multiplication	31	47	33	43
	Division	21	24	24	28

Note. DD problems have two decimal operands; WD problems have one whole number and one decimal as operands.

Analysis of individual children's direction of effect judgments yielded findings consistent with this interpretation. The number of students accurate on 100% of the WD problems was very similar to the number of participants accurate on 100% of the DD problems in every combination of arithmetic operation and decimal size, except for multiplication problems with decimals below 1. On multiplication problems with decimals below 1, almost twice as many children were correct on all three WD problems as on all three DD problems (39% vs. 22% of the sample).

Judgment-plus-explanation task. Comparing the leftmost two columns with the rightmost two columns of Table 2 indicated that judgment accuracy was very similar when explanations were and were not sought. Therefore, analyses of the judgment-plus-explanation task focus on the explanations. All explanations were classified independently by two raters. Percent agreement was 91% (Cohen's kappa was 0.85, above the adequate value of 0.75; Fleiss, 1981). Discussion between the raters was used to resolve discrepancies.

Most explanations (89%) fell into one of three categories:

1. Operation-and-operand explanations (14% of trials). Statements referring to both the operation and the operands or type of operands: "Multiplying with very small decimals makes the value of larger numbers go down;" "If you are multiplying by a number less than 1, you will get a lower outcome."

2. Unconditional operation explanations (56%). Statements about an operation without reference to the operands or type of operand. This category includes rules such as "Multiplication makes bigger" and "When you divide, the number decreases." Also included in this category are statements that implicitly assume that the effect of an operation is the same regardless of the type of operands (e.g., "9.74 * 5.495 will be greater than 9.74 because it's multiplication").

3. Computational estimation explanations (19%). Statements based on rounding of operands and approximate computation (e.g., for 9.74 * 5.495 > 9.74: "Greater because 9 * 5 is 45, which is greater than 9.74").

The remaining explanations were labeled "uninformative" (11%). Of these, 8% could not be categorized (e.g., "because I know" or "you are making the number smaller"), and 3% where the child did not advance an explanation or the explanation was lost.

Frequency of each type of explanation varied with features of the problems. We examined these relations separately for each type of explanation.

Operation-and-operand explanations. Frequency of operation-and-operand explanations varied with the operation, $\chi^2(2, 648) = 15.45$, $p < .01$. It was less common on addition (6% of trials) than on multiplication (19%; $\chi^2(1, 432) = 15.68$, $p < .01$) and division (14%; $\chi^2(1, 432) = 7.46$, $p < .01$). The difference is consistent with the fact that operand size is irrelevant to the direction of effect for addition of positive numbers, but it does influence direction of effect for multiplication and division, making citation of operand size relevant for them.

Frequency of operation-and-operand explanations also varied with the size of the operands, but only on multiplication problems. Such explanations were more common on multiplication problems with decimals below than above 1 (25% vs. 12%; $\chi^2(1, 216) = 6.01$, $p = .01$). Frequency of operation-and-operand explanations did not differ significantly between DD (10%) and WD (15%) problems.

Unconditional operation explanations. Frequency of unconditional operation explanations varied with the arithmetic operation, $\chi^2(2, 648) = 15.45$, $p < .01$. They were more common on addition (60% of trials) and division (61%) than on multiplication (50%).

Computational estimation explanations. Frequency of computational estimation explanations varied with the arithmetic operation, $\chi^2(2, 648) = 20.74$, $p < .01$. They were less frequent with division (10% of trials) than with multiplication (19% of trials; $\chi^2(1, 432) = 6.71$, $p = .01$) and addition (27% of trials; $\chi^2(1, 432) = 20.80$, $p < .01$). Lower frequency of computational estimation on division problems is consistent with it being less well understood than the other arithmetic operations (Carey, 2011; Foley & Cawley, 2003).

Computational estimation explanations were also more common on problems with decimals above 1 than below 1, but only for multiplication (32% vs. 7% of trials; $\chi^2(1, 216) = 21.95$, $p = .01$) and division (15% vs. 6% of trials; $\chi^2(1, 216) = 5.06$, $p = .02$). The whole number part of the operands seemed to facilitate computational estimation on multiplication and division problems by allowing answers based solely on multiplying or dividing the whole number components.

Relations of explanations to direction of effect judgments. Type of explanation was strongly associated with accuracy of direction of effect judgments on multiplication and division problems with decimals below 1 (see Table 3). This relation was only meaningful on these two types of problems, because accuracy was near ceiling for direction of effect judgments on problems with all other combinations of operation and operand size.

As shown in Table 3, operation-and-operand explanations were associated with high accuracy on both multiplication and division problems with decimals below 1. Despite this type of explanation being stated on only 26% of multiplication and 16% of division trials with operands below 1, it was advanced on 65% of trials with correct multiplication judgments and 54% of trials with correct division judgments. These explanations probably reflect students grappling with how to integrate what they know about multiplication and division in general with what they know about results of those operations with numbers from 0 to 1.

Table 3
Percent Correct Direction of Effect Judgments on Multiplication and Division Items With Decimals Below 1 Associated With Each Explanation of Reasoning

Type of explanation	Multiplication of decimals below 1	Division of decimals below 1
Operation and operand	93	88
Unconditional operation	5	2
Computational estimation	44	38
Unspecified	44	45

In contrast, unconditional operation explanations were associated with very low accuracy on both multiplication and division problems with operands below 1, less than 10% correct. In the context of these problems, citing the operation but not the operands, probably reflected the assumption that operand size is irrelevant to the direction of effect, as it is in adding and subtracting positive numbers and in multiplying and dividing numbers above 1.

On these multiplication and division problems with decimals below 1, explanations based on computational estimation were associated with low accuracy, though not as low as with unconditional operation explanations. One reason for the relatively low accuracy was that the two and three digit decimals in the problems made computational estimation difficult unless children rounded the decimals appropriately, which many did not. Another reason was that even when children were correct on the arithmetic, they often transformed answers they obtained so that they were consistent with their general assumption that multiplication yields answers larger than the operands, and division yields answers smaller than the number being divided. One child's explanations for the problems $0.87 * 0.291$ and $0.87 \div 0.291$ illustrates these difficulties. On the multiplication problem, the child said, "If you multiply 0.87 and 0.291, your answer comes to be around 2.793. $2.793 > 0.87$." On the division problem, the child explained: "If you divide 87 by 29 you end up with 3 leaving you with $0.31 < 0.87$."

Repeated addition explanations. Contrary to our expectation, none of the students' explanations referred to solving WD multiplication judgment problems with a decimal below 1 by using repeated addition—estimating the result of adding the decimal the number of times indicated by the whole number (e.g., $5 * 0.291$ interpreted as five iterations of 0.291). In contrast, many explanations were compatible with an unanticipated type of part-whole logic, in which the whole number in the WD problem is the whole and the decimal indicates multiplication by a number that is part of the unit "one" (e.g., "You are multiplying five by a number less than one so the solution is going to be less than one whole five;" "You are multiplying a number by a decimal, and that will make the number go down;" "You're losing stuff when you multiply by a decimal"). These were classified as "operation-and-operand" explanations in the overall categorization of explanations, but this subset of the category seemed worth separate consideration.

Consistent with these examples, on literally all WD multiplication problems with a decimal below 1 in which an explanation associated with a correct judgment treated the two operands asymmetrically, the decimal was treated as the operator and the whole number as the object of the operation. This approach was observed on 26% of WD multiplication problems. (Interrater agreement in coding this type of part-whole explanation was 93% (Cohen's kappa was = 0.77.)

Magnitude comparison. Students correctly answered 88% of the decimal magnitude comparisons. Most students (76%) were accurate on more than 95% of the decimal comparisons; 9% of students consistently ignored the decimal points in the numbers being compared.

Number of correct decimal magnitude comparison and direction of effect judgments were weakly related. On the judgments only task, the relation was significant, $r = .35$, $p < .05$; on the judgments plus explanations condition, it was not ($r = .26$, *ns*).

General Discussion

This study extended prior ones in examining direction of effect judgments with decimals rather than fractions, problems involving a whole number and a rational number as well as two rational numbers, and measures that included confidence ratings and explanations of direction of effect judgments. Each of these features clarified the meaning of direction of effect judgments, sometimes in ways that differed from our expectations, and suggested means for improving instruction to increase students' understanding of rational number arithmetic.

One clear finding was that inaccurate direction of effect judgments for multiplication and division of fractions are not attributable only to difficulty understanding fraction notation. Identical difficulties were present with decimals, a notation that maps more transparently onto whole number notation. Thus, lack of understanding of the direction of effect of multiplying and dividing numbers below 1 is general to positive rational numbers, rather than being specific to fractions. Minimal correlations between accuracy of direction of effect judgments and accuracy on both magnitude comparison and arithmetic computation added evidence that this lack of understanding could not be attributed to lack of either magnitude or arithmetic knowledge.

Confidence ratings indicated differences between two groups of children. The half of children whose direction of effect judgments for decimal arithmetic invariably matched the pattern for the corresponding whole number arithmetic operation were highly confident in their incorrect judgments regarding multiplication and division of decimals below 1. Their confidence in these incorrect judgments was not only very high in absolute terms, it was as high as their confidence in their correct judgments of the direction of effect of addition, subtraction, and multiplication of operands above 1. Thus, these children's performance matched the strong conviction interpretation of direction of effect judgments.

In contrast, the half of children whose judgments less consistently matched the whole number pattern were less confident in some of their judgments. This was particularly the case for the older children (8th graders) who were less confident in their direction of effect judgments involving decimals below 1, especially on multiplication and division problems. This was consistent with the cognitive conflict interpretation. This finding might reflect the 8th graders whose judgments were less consistent beginning to suspect that the direction of effect of multiplication and division with numbers from 0 to 1 differs from that with operands above 1, but remaining uncertain. Examination of high school students' fraction and decimal direction of effect judgments and their confidence in those judgments could indicate whether understanding, or at least uncertainty, continues to grow with further mathematical experience.

The explanations data revealed a new phenomenon and improved understanding of another. The new phenomenon was that for both multiplication and division of decimals below 1, direction of effect judgments vary greatly with the type of explanation that children generate. Explanations that noted both the arithmetic operation and whether the operands were above or below 1 were strongly associated with correct judgments; 90% of judgments that preceded such explanations were accurate. In contrast, less than 50% of judgments were correct when explanations cited only the type of operation, indicated reliance on computational estimation, or did not indicate any basis for the judgment. These data are consistent with the view that encoding not only the type of operation but also whether the operands are above or below 1 is essential to understanding rational number arithmetic.

The explanations data also changed our understanding of the finding that students were more accurate when judging the direction of effect on multiplication problems that involve a whole number and a decimal below 1 than when making such judgments on multiplication problems with two decimals below 1. This effect was quite consistent; judgments were more accurate on all multiplication problems that included a whole number and a decimal below 1 than on any problem that included two decimals below 1.

Although we predicted this finding, the explanations data revealed that our prediction was right for a wrong reason. The explanations showed no evidence for the hypothesized reliance on the logic of repeated addition to solve multiplication problems that involved a whole number and a decimal below 1. Instead, most explanations that accompanied correct direction of effect judgments on such problems relied on a kind of part-whole logic. That is, the explanations emphasized that multiplying a whole number by a decimal less than 1 meant taking only a part of the whole number. In other words, rather than viewing the whole number as indicating the number of iterations of the decimal, children viewed the whole number as a whole and reasoned that multiplying by a number less than 1 would leave only part of the whole.

The same logic could have been applied to multiplication of two decimals between 0 and 1—there too, multiplying by a number less than 1 would leave only part of the original number—but it rarely was. One possibility is that greater familiarity with whole numbers might facilitate thinking about the effects of multiplying them by other numbers, perhaps through whole numbers being easier to encode as objects on which other multiplicands might operate. Another, nonexclusive, possibility is that the coincidence between the term *whole number* and that number serving as the whole in this context, promoted this reasoning.

The present research extended previous findings about direction of effect knowledge of decimals in at least three ways. One was demonstrating that similar findings emerge with more focused measures of direction of effect knowledge, judgments of the direction of effect in inequalities, as with the less focused measures of this knowledge used previously (selection of operations in word problems and unsolicited expressions of surprise; Fischbein et al., 1985; Graeber & Tirosh, 1990; Tirosh & Graeber, 1989). Another extension involved demonstrating that observations with fractions in these and our own previous study were not unique to fractions; rather, they extend to decimals as well. Third, the present findings narrowed the range of alternative explanations of the inaccurate judgments by showing that inaccurate direction of effect judgments were not due only to weak knowledge of operand magnitudes or computational procedures. Inaccurate direction of effect judgments with multiplication and division of decimals between 0 and 1 was observed even though most participants exhibited excellent understanding of decimal magnitudes and arithmetic procedures.

The findings also raise an intriguing theoretical question. Theories of error learning (e.g., Ohlsson, 1996; Ohlsson & Rees, 1991) propose that when people detect errors, they narrow their generalizations and subsequently err less often. The high frequency of direction of effect errors in the present study raises the issue of

why such errors remain so frequent after years of fraction arithmetic experience. Do learners not notice the pattern that multiplying two numbers between 0 and 1 always yields an answer smaller than either multiplicand? Do teachers not point out the pattern? Do children stop trying to make sense of rational number arithmetic, and therefore solely focus on executing procedures correctly rather than trying to identify relations between problems and answers? Specifying why these errors persist for so long, despite learners' substantial experience with rational number arithmetic, may prove useful in elaborating theories of error learning so that they can predict not only learning but also failures to learn.

Implications for Instruction

A general instructional implication of the present findings, especially taken together with the parallel findings of Siegler and Lortie-Forgues (2015) with fractions, is that at least some goals of the Common Core State Standards regarding understanding of rational number arithmetic are not yet being attained. For instance, interpreting multiplication as scaling (i.e., scaling up when multiplying by a number above 1 and scaling down when multiplying by a number below 1) is one of the main learning goals of the Common Core (CCSSI, 2015) for 5th graders. If students had such understanding, they would have been much more accurate on the direction of effect task with both decimals and fractions than they turned out to be. To the extent that these findings are general, they suggest that current approaches to teaching conceptual understanding of rational number arithmetic need to be improved.

A more specific instructional implication was suggested by our finding that children were more accurate when multiplying a whole number by a decimal between 0 and 1 than when multiplying two decimals of that size. This finding suggests that focusing on the former type of problem provides a useful transition between whole number multiplication and multiplication of two rational numbers. The fact that the part–whole logic was seen less often on multiplication problems with two decimals below 1, despite being equally applicable to both types of problems, suggests that substantial transfer of the reasoning to such problems requires specific efforts to promote it. The instructional implication is that learning would benefit from teachers and textbooks presenting well-chosen analogies that highlight that the same reasoning applies to DD as to WD problems. Instruction based on structurally sound analogies has often proved effective in improving numerical understanding (e.g., Chen, Lu, & Holyoak, 2014; Opfer & Siegler, 2007; Sullivan & Barner, 2014). The clear parallels between multiplication of a whole number and a rational number between 0 and 1, and two rational numbers between 0 and 1, suggest that promoting analogies from the easier to the harder case could improve learning.

Another implication is that instruction should explicitly challenge students' belief that arithmetic with all numbers consistently works like arithmetic with whole numbers. Children whose direction of effect judgments invariably followed the whole number pattern were highly confident in the correctness of incorrect as well as correct judgments. Confidence is often a good thing, but misplaced confidence is not. One way to challenge the mistaken belief would be to focus students' attention on contradictory evidence. Students could predict the direction of effect of multiplication of rational numbers below 1, and then compare their judgment with the actual answer generated by their own computation. Teachers could complement this activity with questions about why answers were wrong, as apparent contradictions alone could be ignored or attributed to calculation errors (Vosniadou, Ioannides, Dimitrakopoulou, & Papademetriou, 2001). Confronting students with contradictory evidence is a common and effective teaching practice in other domains where misconceptions are frequent, such as science education (e.g., Chinn & Brewer, 1993). Moreover, people with high confidence in their errors have been found to be particularly responsive to feedback contradicting their beliefs (e.g., Butterfield & Metcalfe, 2001).

A further instructional implication is that students should be encouraged to consider both the size of the operands and the arithmetic operation when judging direction of effect of arithmetic operations. Explanations that cited both variables consistently accompanied correct judgments. By contrast, explanations that only cited the type of operation almost always accompanied incorrect judgments. Juxtaposing problems that involve operands below 1 with problems that involve operands above 1, and asking students to reflect about why they need to consider the size of the operand as well as the operation, might prove effective at raising students' awareness of the relevance of both the operation and the operands to direction of effect judgments for multiplication. It might also help to increase their understanding of multiplication more generally.

Limitations and Future Directions

The present study has several limitations, each of which suggests directions for future research. One limitation is that our study does not address the effects of variations in mathematics curricula. Students who received more conceptually oriented instruction might show greater understanding of the direction of effect of rational number arithmetic operations. In a similar vein, the more accurate judgments on WD problems than on DD problems in Study 2 might reflect children encountering WD problems more often; without detailed knowledge of the input that children received, it was impossible to evaluate this interpretation, but the effects of curricula and instructional input more generally should be evaluated in future research.

Another limitation is that the present study did not directly compare direction of effect knowledge with decimals and fractions, and thus did not address the possibility that the notation moderates the strength of the observed effects. Future studies could test this possibility by presenting both fraction and decimal direction of effect problems to the same participants.

The present research also could not specify the role of teacher and textbook input on students' direction of effect knowledge. We attempted to contact the two teachers who taught the children in the study. One teacher indicated that she did not use a textbook but rather a variety of materials gathered from the Internet; we could not locate the other teacher, who had left the school by the time we attempted to address this issue. The superior performance on WD relative to DD multiplication problems with operands between 0 and 1 might have been due to students encountering more WD than DD problems, or it might have been due to WD problems more often being presented with aids to conceptual understanding, such as manipulatives or number lines. In the absence of detailed data on the input that students received, this hypothesis could not be tested in the present study.

A further limitation of the present study is that idiosyncratic features of the task might have influenced students' reasoning. For instance, to allow identical operand orders for all four arithmetic operations without requiring understanding of negatives, we always presented the larger operand first and used it as the comparison answer (e.g., $5 * .291 > 5$). This ordering, and the consequence of always having the whole number as the first operand on WD problems, might have influenced students' reasoning. In particular, presenting problems in which the whole number operand was second, such as $0.291 * 5 > 0.291$, might have focused students' attention on the changes to 0.291 caused by being multiplied by 5 and thus led them to see the problem in terms of repeated addition. Another possibility is that phrasing the questions differently (e.g., "If you calculate how much 5 of the 0.291's is, will the answer be greater than 5?") might have revealed greater use of the repeated addition approach than the format used here (e.g., "Is $5 * 0.291 > 5$?"). Testing the effects of these and other features of the direction of effect procedure would be valuable for evaluating the generality of the conclusions yielded by this study, as well as for suggesting ways of improving children's conceptual understanding of rational number arithmetic.

References

Bailey, D. H., Hoard, M. K., Nugent, L., & Geary, D. C. (2012). Competence with fractions predicts gains in mathematics achievement. *Journal of Experimental Child Psychology, 113,* 447–455. http://dx.doi.org/10.1016/j.jecp.2012.06.004

Bell, M., Bretzlauf, J., Dillard, A., Hartfield, R., Isaacs, A., McBride, J., . . . Saecker, P. (2007). *Everyday mathematics sixth grade math* (3rd ed.). Chicago, IL: McGraw-Hill.

Booth, J. L., Newton, K. J., & Twiss-Garrity, L. K. (2014). The impact of fraction magnitude knowledge on algebra performance and learning. *Journal of Experimental Child Psychology, 118,* 110–118. http://dx.doi.org/10.1016/j.jecp.2013.09.001

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1491–1494. http://dx.doi.org/10.1037/0278-7393.27.6.1491

Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology, 27,* 777–786. http://dx.doi.org/10.1037/0012-1649.27.5.777

Carey, S. (2011). *The origin of concepts*. New York, NY: Oxford University Press.

Carpenter, T., Corbitt, M., Kepner, H., Lindquist, M., & Reys, R. (1980). Results of the second NAEP mathematics assessment: Secondary school. *Mathematics Teacher, 73,* 329–338.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction.* Portsmouth, NH: Heineman.

Carpenter, T., Lindquist, M., Matthews, W., & Silver, E. (1983). Results of the third NAEP mathematics assessment: Secondary school. *Mathematics Teacher, 76,* 652–659.

Charles, R. I., Illingworth, M., McNemar, B., Mills, D., & Ramirez, A. (2012). *Prentice Hall mathematics: Courses 2 common core edition.* New York, NY: Pearson.

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology, 71,* 27–54. http://dx.doi.org/10.1016/j.cogpsych.2014.01.002

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63,* 1–49. http://dx.doi.org/10.3102/00346543063001001

College Board. (2014). *AP chemistry course and exam description* [PDF document]. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-chemistry-course-and-exam-description.pdf

College Board. (2015). *Advanced Placement Physics 1 equations, effective 2015* [PDF document]. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-physics-1-equations-table.pdf

Common Core State Standards Initiative (CCSSI). (2015). *Common core state standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers. Retrieved from http://www.corestandards.org/Math/

Desmet, L., Gregoire, J., & Mussolin, C. (2010). Developmental changes in the comparison of decimal fractions. *Learning and Instruction, 20,* 521–532. http://dx.doi.org/10.1016/j.learninstruc.2009.07.004

DeWolf, M., Grounds, M. A., Bassok, M., & Holyoak, K. J. (2014). Magnitude comparison with different types of rational numbers. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 71–82. http://dx.doi.org/10.1037/a0032916

English, L., & Halford, G. (1995). *Mathematics education: Models and processes*. Mahwah, NJ: Erlbaum.

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology, 123,* 53–72. http://dx.doi.org/10.1016/j.jecp.2014.01.013

Fischbein, E., Deri, M., Nello, M. S., & Marino, M. S. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education, 16,* 3–17. http://dx.doi.org/10.2307/748969

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: Wiley.

Foley, T., & Cawley, J. (2003). About the mathematics of division: Implications for students with learning disabilities. *Exceptionality, 11,* 131–149. http://dx.doi.org/10.1207/S15327035EX1103_02

Graeber, A. O., & Tirosh, D. (1990). Insights fourth and fifth graders bring to multiplication and division with decimals. *Educational Studies in Mathematics, 21,* 565–588. http://dx.doi.org/10.1007/BF00315945

Hecht, S. A. (1998). Toward an information-processing account of individual differences in fraction skills. *Journal of Educational Psychology, 90,* 545–559. http://dx.doi.org/10.1037/0022-0663.90.3.545

Hecht, S. A., & Vagi, K. J. (2010). Sources of group and individual differences in emerging fraction skills. *Journal of Educational Psychology, 102,* 843–859. http://dx.doi.org/10.1037/a0019824

Hiebert, J., & Wearne, D. (1985). A model of students' decimal computation procedures. *Cognition and Instruction, 2,* 175–205. http://dx.doi.org/10.1080/07370008.1985.9648916

Hiebert, J., & Wearne, D. (1986). Procedures over concepts: The acquisition of decimal number knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 199–223). Hillsdale, NJ: Erlbaum.

Hoyles, C., Noss, R., & Pozzi, S. (2001). Proportional reasoning in nursing practice. *Journal for Research in Mathematics Education, 32,* 4–27. http://dx.doi.org/10.2307/749619

Iuculano, T., & Butterworth, B. (2011). Understanding the real value of fractions and decimals. *Quarterly Journal of Experimental Psychology, 64,* 2088–2098. http://dx.doi.org/10.1080/17470218.2011.604785

Laski, E. V., & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development, 78,* 1723–1743. http://dx.doi.org/10.1111/j.1467-8624.2007.01087.x

Lortie-Forgues, H., Tian, J., & Siegler, R. S. (2015). Why is learning fraction and decimal arithmetic so difficult? *Developmental Review, 38,* 201–221. http://dx.doi.org/10.1016/j.dr.2015.07.008

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues*

(NCEE 2009–013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Mazzocco, M. M., & Devlin, K. T. (2008). Parts and "holes": Gaps in rational number sense among children with vs. without mathematical learning disabilities. *Developmental Science, 11,* 681–691. http://dx.doi.org/10.1111/j.1467-7687.2008.00717.x

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Ni, Y., & Zhou, Y.-D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist, 40,* 27–52. http://dx.doi.org/10.1207/s15326985ep4001_3

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103,* 241–262. http://dx.doi.org/10.1037/0033-295X.103.2.241

Ohlsson, S., & Rees, E. (1991). The function of conceptual understanding in the learning of arithmetic procedures. *Cognition and Instruction, 8,* 103–179. http://dx.doi.org/10.1207/s1532690xci0802_1

Ojose, B. (2015). *Common misconceptions in mathematics: Strategies to correct them*. Lanham, MD: University Press of America.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55,* 169–195. http://dx.doi.org/10.1016/j.cogpsych.2006.09.002

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education, 20,* 8–27. http://dx.doi.org/10.2307/749095

Resnick, L. B., & Omanson, S. F. (1987). Learning to understand arithmetic. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 3, pp. 41–95). Hillsdale, NJ: Erlbaum.

Rizvi, N. F., & Lawson, M. J. (2007). Prospective teachers' knowledge: Concept of division. *International Education Journal, 8,* 377–392.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26,* 521–562. http://dx.doi.org/10.1207/s15516709cog2605_1

Sformo, T. (2008). *Practical problems in mathematics: For automotive technicians*. Independence, KY: Cengage Learning.

Siegler, R. S. (1987). The perils of averaging over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264. http://dx.doi.org/10.1037/0096-3445.116.3.250

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., . . . Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23,* 691–697. http://dx.doi.org/10.1177/0956797612440101

Siegler, R. S., & Lortie-Forgues, H. (2015). Conceptual knowledge of fraction arithmetic. *Journal of Educational Psychology, 107,* 909–918. http://dx.doi.org/10.1037/edu0000025

Siegler, R. S., & Pyke, A. A. (2013). Developmental and individual differences in understanding of fractions. *Developmental Psychology, 49,* 1994–2004. http://dx.doi.org/10.1037/a0031200

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62,* 273–296. http://dx.doi.org/10.1016/j.cogpsych.2011.03.001

Sullivan, J., & Barner, D. (2014). The development of structural analogy in number-line estimation. *Journal of Experimental Psychology, 128,* 171–189. http://dx.doi.org/10.1016/j.jecp.2014.07.004

Tirosh, D., & Graeber, A. O. (1989). Preservice elementary teachers' explicit beliefs about multiplication and division. *Educational Studies in Mathematics, 20,* 79–96. http://dx.doi.org/10.1007/BF00356042

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction, 11,* 381–419. http://dx.doi.org/10.1016/S0959-4752(00)00038-4

Received December 9, 2015
Revision received July 11, 2016
Accepted July 12, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 3, 387–404

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000146

Making Connections: Replicating and Extending the Utility Value Intervention in the Classroom

Chris S. Hulleman and Jeff J. Kosovich
University of Virginia

Kenneth E. Barron and David B. Daniel
James Madison University

We replicated and extended prior research investigating a theoretically guided intervention based on expectancy-value theory designed to enhance student learning outcomes (e.g., Hulleman & Harackiewicz, 2009). First, we replicated prior work by demonstrating that the utility value intervention, which manipulated whether students made connections between the course material and their lives, increased both interest and performance of low-performing students in a college general education course. Second, we extended prior research by both measuring and manipulating one possible pathway of intervention effects: the frequency with which students make connections between the material and their lives. In Study 1, we measured connection frequency and found that making more connections was positively related to expecting to do well in the course, valuing the course material, and continuing interest. In Study 2, we manipulated connection frequency by developing an enhanced utility value intervention designed to increase the frequency with which students made connections. The results indicated that students randomly assigned to either utility value intervention, compared with the control condition, subsequently became more confident that they could learn the material, which led to increased course performance. The utility value interventions were particularly effective for the lowest-performing students. Compared with those in the control condition who showed a steady decline in performance across the semester, low-performing male students randomly assigned to the utility value conditions increased their performance across the semester. The difference between the utility value and control conditions for low-performing male students was strongest on the final exam (d = .76).

Keywords: academic motivation, educational intervention, expectancy-value motivation, gender, utility value

Supplemental materials: http://dx.doi.org/10.1037/edu0000146.supp

Optimizing student motivation and learning in the classroom is a goal shared by most educators. However, there is no consensus on the best methods. Rewarding students for classroom behavior or performance, or threatening punishment, are strategies conventionally believed to increase motivation and engagement (Ash, 2008; Kohn, 1999; Newby, 1991). Such strategies presume that learning tasks are not inherently rewarding and, therefore, extrinsic reasons for task engagement must be introduced. In contrast, tapping into more intrinsic sources of motivation (Ames, 1992), such as fostering individual interest in specific topics (Hidi & Renninger, 2006), self-determined motivation (Deci & Ryan, 1985), and self-directed task involvement (Csikszentmihalyi, 1990), are strategies more likely to be recommended by educational psychologists (Boekaerts, 2002). By focusing on student perceptions and beliefs about the value of the learning activity, contemporary models of expectancy-value motivation highlight this more intrinsic source of motivation (e.g., Brophy, 1999; Eccles et al., 1983). Although the research generated by the expectancy-value framework has been largely correlational (Wigfield & Cambria, 2010), recent classroom studies reveal that interventions designed to enhance perceptions of value can increase both interest and course performance (e.g., Hulleman & Harackiewicz, 2009; Hulleman, Godes, Hendricks, & Harackiewicz, 2010). The research presented herein replicates and extends this prior work by further investigating a theory-based intervention designed to enhance student motivation and performance.

The Expectancy-Value Framework

Originally adapted from classic models of expectancy-value motivation (e.g., Atkinson, 1957; Vroom, 1964), Eccles and her colleagues (1983) proposed that motivation in educational contexts

This article was published Online First August 15, 2016.

Chris S. Hulleman, Center for Advanced Study of Teaching and Learning, University of Virginia; Jeff J. Kosovich, Curry School of Education, University of Virginia; Kenneth E. Barron and David B. Daniel, Department of Psychology, James Madison University.

This research was supported by the National Science Foundation, through Grants DRL 1252463 and 1228661, to the first author, and by the U.S. Department of Education, through Grant R305B090002, to the second author. The opinions expressed are those of the authors and do not represent views of the funding agencies. A previous version of this article was presented at the annual conference of the Northeastern Educational Research Association, October 2012.

Correspondence concerning this article should be addressed to Chris S. Hulleman, Center for the Advanced Study of Teaching and Learning, University of Virginia, 405 Emmet Street South, Charlottesville, VA 22903. E-mail: chris.hulleman@virginia.edu

is determined most proximally by an individual's expectancy beliefs and subjective task values. Expectancy beliefs are defined as the belief that one can succeed at an activity, and have been correlated with achievement outcomes and achievement choices, such as continued persistence and course taking (for reviews see Richardson, Abraham, & Bond, 2012; Robbins et al., 2004). Subjective task values are defined as the perceived importance of a task or activity, and four facets were originally proposed by Eccles and colleagues (1983): intrinsic (enjoyment), utility (usefulness for proximal or distal goals), attainment (importance for one's sense of self), and cost (psychological barriers to, and negative consequences of, task engagement). A wealth of prior research has demonstrated that task values are positively correlated with continued persistence and ongoing motivation in an activity (for reviews see Wigfield & Cambria, 2010), except cost which is negatively related (e.g., Conley, 2012; Flake, Barron, Hulleman, McCoach, & Welsh, 2015). Consistent with more recent conceptualizations of the expectancy-value framework (e.g., Barron & Hulleman, 2015), we consider cost to be a unique construct independent of expectancy and value.

In particular, students' perceptions of utility value have been associated with achievement outcomes in longitudinal field studies (e.g., Bong, 2001; Durik, Vida, & Eccles, 2006; Hulleman et al., 2008). Originally defined as "the value a task acquires because it is instrumental in reaching a variety of long- and short-range goals (Eccles & Wigfield, 1995, p. 216)," measures of utility value have captured the relationship between students' current (e.g., classes, hobbies) and future goals (e.g., college major, career). For example, one of the original scales measuring utility value (1995) included items that tapped students' future plans ("How useful is learning advanced high school math for what you want to do after graduation?") and current goals ("How useful is what you learn in advanced high school math for your daily life outside school?"). Recent measures of utility value have mirrored this connection to both current and future goals (e.g., Hulleman et al., 2008). Furthermore, because some goals are more personally important than others, utility value has been conceptualized as having elements of both intrinsic and extrinsic motivation (Hulleman et al., 2008; Simons, Vansteenkiste, Lens, & Lacante, 2004).

Despite extensive correlational support, limited research has tested the effectiveness of interventions based on expectancy-value models. Our review of the current literature revealed only a handful of published papers investigating interventions based on the expectancy-value framework in an educational context, all focused on utility value (Acee & Weinstein, 2010; Brown, Smith, Thoman, Allen, & Muragishi, 2015; Harackiewicz, Canning, Tibbetts, Priniski, & Hyde, 2015; Harackiewicz, Rozek, Hulleman, & Hyde, 2012; Hulleman & Harackiewicz, 2009; Hulleman et al., 2010; Johnson & Sinatra, 2013). To provide stronger claims about both internal and external validity, three of the studies were conducted as double-blind, randomized classroom experiments (Harackiewicz et al., 2015; Hulleman & Harackiewicz, 2009; Hulleman et al., 2010, Study 2). Hulleman and colleagues evaluated a utility value intervention that encouraged students to discover the relevance of the material they were studying to their lives. Utility value was manipulated through a writing prompt given to high school science (Hulleman & Harackiewicz, 2009) and college psychology (Hulleman et al., 2010) students as part of their regularly assigned coursework. Students were randomly assigned to either write about the relevance and usefulness of the course material in their own lives (relevance condition) or a summary of the material they were currently studying (control condition). In the high school sample, students completed writing assignments every three to four weeks of a 20-week semester. Students averaged about five essays throughout the semester. In the college sample, students were given writing assignments in the 8th and 12th weeks of a 15-week semester. The key dependent variables in both studies were end-of-semester interest in the course topic and course grades. The researchers provided teachers with information regarding whether students had completed the essays, but teachers were blind to condition throughout the semester. Because students wrote about course-related topics in both the relevance and control conditions, knowledge activation was controlled (i.e., summarization; see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). The conditions thereby differed only in terms of the activation of utility value.

In these studies, the results indicated that the intervention was more effective for students with lower perceived or actual competence. In college psychology, students who performed more poorly on initial course exams were more interested in the course if they were in the relevance condition than the control condition. In high school science, the interaction effect was replicated on both science interest and grades for students who entered the course with lower performance expectations. In fact, the effect on end-of-semester GPA for students with low performance expectations resulted in an increase in .80 GPA points. In both studies, the intervention increased students' perceptions of utility value, and these increased perceptions led to improved performance and interest. Furthermore, this pattern of intervention effectiveness was also replicated with (a) undergraduate students who learned a mental math technique in the laboratory (Hulleman et al., 2010, Study 1), (b) first-generation college students enrolled in introductory science classes (Harackiewicz et al., 2015), and (c) high school students whose parents received an intervention on how to talk to their teenager about the value of math and science coursework (Rozek, Hyde, Svoboda, Hulleman, & Harackiewicz, 2015).

Learning Why the Utility Value Intervention Works

Together, these initial studies demonstrated that interventions designed to increase subjective task value subsequently increased interest and performance. However, these studies also demonstrated that the intervention effects were not the same for everyone (for reviews see Durik, Hulleman, & Harackiewicz, 2015; Harackiewicz, Tibbetts, Canning, & Hyde, 2014). Given these intervention effects, we sought to understand what might explain the underlying mechanisms, as well as develop future interventions that might work for all students regardless of their success expectancies and performance history.

Although Hulleman and colleagues (2009, 2010; Harackiewicz et al., 2015) have generally used relevance and utility value interchangeably in describing their intervention, there is a potentially important distinction to be made. Whereas utility value refers to usefulness to a proximal or distal goal, *relevance* simply refers to the presence of a relationship between one topic or idea and another topic or idea, which could include a goal but also includes a broader set of relationships (Kosovich & Hulleman, 2016). For example, math could be useful because it will help me in a future

job (*utility value*), or it could relate to my life because store cashiers need it even if I do not (*relevance*). Because the utility value intervention is one type of relevance intervention, one possible mechanism for utility value intervention effects is that encouraging students to find a connection allows them to notice relationships that they previously had not. Seeing such connections may allow individuals to view new information from a different perspective, and develop a more in-depth integration of their knowledge (Bransford & Schwartz, 1999). In addition, simply referencing the self when learning new material can lead to learning gains (e.g., Barney, 2007; for a review see Symons & Johnson, 1997). Consistent with this hypothesis is the finding that instructing individuals to find connections between learning situations can increase the likelihood of adapting a skill from one situation to another (i.e., cognitive transfer; Burke & Hutchins, 2007; Gentner, Loewenstein, & Thompson, 2003; Gick & Holyoak, 1980). As hypothesized in the early work of learning theorists (e.g., Thorndike & Woodworth, 1901), making a connection may enhance learning by instigating a set of processes that engenders a different approach to studying that may increase learning. For example, if a student finds a personal connection during a psychology lecture, the student may be more interested in the assigned reading and to discuss the material with friends. In general, the student may be more motivated to actively process the material during lecture and later when reading the book. Establishing relationships between new knowledge and old ideas may create a richer cognitive architecture which the student can draw upon when studying. As a result, students who make more connections between course material and existing knowledge may be more likely to find usefulness in the course, which may enhance motivation.

What is the best way to investigate the frequency of connections as a key pathway through which the utility value intervention impacts outcomes? One approach is to measure the proposed mechanism and conduct path analyses (e.g., Hulleman et al., 2010). This approach is appealing because it is relatively simple and falls within the range of most statistical packages (e.g., Tofighi & MacKinnon, 2011). The limitation of this approach is that it is correlational, and it does not account for other key variables that may explain the effects of the intervention but have not been measured. In contrast, the second approach, which is far less common but more powerful, is to manipulate the mechanism (see Baron & Kenny, 1986; Sigall & Mills, 1998). This approach allows the researcher to randomly assign participants to different levels of the variable to establish a cause-and-effect relationship. The con to this approach, which is inherent to all intervention studies, is that the effect of a manipulated variable may not be the same as the effect of the measured variable (cf. Barron & Harackiewicz, 2001). Rather than choosing one method, both approaches to enhancing learning outcomes will be investigated in this paper.

Interest as an Educational Outcome

Academic performance is a widely accepted educational outcome and grades play a pivotal role in a student's long-term educational opportunities. A less-acknowledged but equally important outcome is interest (Hidi & Harackiewicz, 2000; Hulleman et al., 2008). In a longitudinal study, Harackiewicz, Barron, Tauer, and Elliot (2002) found that interest predicted course choice and college major selection over six years, whereas prior performance and college GPA did not. Interest can be thought of as two different types (Hidi & Renninger, 2006). *Situational interest* is the experience of engagement or attention during a task (Schraw & Lehman, 2001). *Individual interest* is an enduring proclivity for the task or behavior (Renninger & Wozniak, 1985). In the current study, we focused on situational interest because it is a precursor for the development of individual interest (Hidi & Renninger, 2006), which predicts long-term academic and career choices (e.g., Peters & Daly, 2013; Pike & Dunne, 2011), and is heavily influenced by the learning context and therefore amenable to change via short-term interventions (e.g., Durik & Harackiewicz, 2007; Hulleman et al., 2010).

Current Studies

In Study 1, students' perceptions of how often they connected the material to their lives was measured and used to predict student learning outcomes over the course of a semester. In Study 2, connection frequency was manipulated through an experimental intervention delivered as part of course embedded assignments. Although we hypothesized that both measured and manipulated connection frequencies will have similar effects on learning outcomes, it is possible that measured and manipulated variables capture different aspects of the same phenomena. Thus, it is crucial to examine both types of effects when investigating the role of connection frequency.

The utility value interventions utilized in the current studies were based on the self-generated utility value interventions used by Hulleman and colleagues (e.g., Hulleman et al., 2010; Hulleman & Harackiewicz, 2009). In addition to replicating this prior research, we extend it in five ways. First, we used an online course management system to deliver the intervention, instead of paper-and-pencil writing assignments used in prior studies, which was seamlessly embedded within the course as a regular course assignment. Second, we tested the mechanism of the utility value intervention by measuring the self-reported frequency with which students made connections between the material and their lives throughout the semester (e.g., connection frequency). Third, we further examined the mechanism of the utility value intervention by manipulating one hypothesized mechanism—connection frequency. Fourth, we examined the effects of the intervention on students' expectancies and perceived costs in the course. The theoretical model hypothesizes that the utility value intervention effects are driven through increased perceptions of value, particularly utility value. However, it is also possible that writing about the relevance of course material could increase students' expectancies that they can learn the material and perform well in the course, or decrease their perceived cost for learning. Fifth, because prior research found differential intervention effects based on key demographics, such as gender and initial performance, we also examined whether the intervention was more effective for students at-risk for poor performance. In the case of a college general education course, students who initially perform poorly in the course are most at-risk, as are male students (Voyer & Voyer, 2014).

Study Samples

The samples for both studies in this paper came from students who were enrolled in two sections of a 15-week introductory psychology course at a midsized university in the southeastern United States. Both sections were taught by the same instructor. Of the 589 students who were enrolled in the course, 501 students (85%) completed the initial consent form and were eligible to participate in our research. Of these students, 113 were randomly selected to participate in Study 1 and 388 were randomly selected for Study 2.

Study 1: A Longitudinal, Correlational Investigation

In Study 1, we explored a potential pathway for utility value effects found in prior research. Specifically, we developed a new measure that asked students to report on the frequency with which they made connections between the material and their lives while listening to lectures, studying for exams, and socializing with friends. As a method for providing initial validity evidence for both the idea and the measure, we examined whether students' self-reports of connection frequency provided a pathway through which utility value was related to student learning outcomes during a semester-long undergraduate psychology course.

Method

Participants. Of 113 eligible introductory psychology students, the final sample included 97 students who were over the age of 18 and participated in the surveys. Students received extra credit in the course for completing both surveys. The sample was 70% female, 84% white (4% African American, 4% Asian), 86% non-psychology majors, and 55% freshman (28% sophomore, 12% junior, 3% senior). The mean age of participants was 18.7 years.

Self-reports of expectancy-value-cost motivation. Students completed self-report surveys at three time points during the semester: Time 1 measures were taken during the 2nd week, Time 2 measures were taken during the 8th week, and Time 3 measures were taken during the 14th week of the semester. Measures of expectancy, utility value, and cost were collected at Time 1 and 3, and have been previously validated with students in middle school (Kosovich, Hulleman, Barron, & Getty, 2015), high school (Hulleman & Harackiewicz, 2009), and college (Grays, 2013; Hulleman et al., 2008). Expectancy was measured using a 4-item scale (e.g., "I expect to do well in this class," α = 92 and .93). Utility Value was measured using a 6-item scale (e.g., "I can apply what we're learning in this class to the real world," "The course material is relevant to my future career plans," α = .93 and .92). Cost was measured using a 6-item scale (e.g., "Doing well in this class isn't worth all the things that I have to give up," α = .81 and .87). All self-report items used an 8-point Likert-type scale that ranged from 1 (*completely disagree*) to 8 (*completely agree*; see Appendix A in the Supplemental Online Material for complete list of items, and Table 1 for descriptive information including reliabilities).

Connection frequency. To capture the number of connections between students' lives and the course material, a 3-item measure of connection frequency was included at Time 2, just after the second course exam (e.g., "When reading a chapter from the textbook/During a regular class period or lecture/When studying for quizzes and exams, how often do you connect the class material to your life?"; α = .87). These items used a 6-point Likert-type scale ranging from 1 (*never*) to 6 (*all of the time*).

Learning outcomes. Two major learning outcomes were collected: academic performance and interest in the course. Students' academic performance was measured using class exam scores. There were 4 noncumulative exams, each covering four chapters in the textbook, and administered during the 4th, 8th, 12th, and 16th weeks of the course. There were 80 multiple-choice questions on the first three exams, and 100 multiple-choice questions on the fourth and final exam. Each question was worth one point. Students completed the exams using Scantron answer sheets and the exams were machine-scored. The grades were never curved and the grading scale was: 90% to 100% A, 80% to 89% B, 70% to

Table 1
Descriptive Statistics for Major Variables in Study 1

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Time 1 interest											
2. Time 1 expectancy	.34										
3. Time 1 utility value	.79	.31									
4. Time 1 cost	−.37	−.20	−.30								
5. Initial exam	.00	.23	.04	−.12							
6. Time 2 connections	.33	.33	.25	−.14	.14						
7. Time 3 interest	.75	.34	.62	−.35	.11	.46					
8. Time 3 expectancy	.33	.60	.32	−.39	.50	.34	.54				
9. Time 3 utility value	.61	.32	.69	−.29	.06	.41	.77	.45			
10. Time 3 cost	−.33	−.13	−.18	.66	−.19	−.30	−.37	−.50	−.29		
11. Final exam	.01	.06	.13	.08	.60	.05	.03	.27	.02	.02	
12. Female	.19	−.03	.18	−.07	.10	−.01	.16	.08	.21	−.08	.20
Observed min.	3.00	4.50	2.33	1.83	31.00	2.00	3.33	4.00	2.00	1.67	35.00
Observed max.	8.00	8.00	8.00	6.00	75.00	6.00	8.00	8.00	8.00	8.00	98.00
Mean	6.00	6.50	6.09	3.60	63.16	3.88	6.14	6.48	6.01	3.96	81.85
SD	1.08	.86	1.11	.93	7.19	.91	1.19	.97	1.15	1.23	9.28
α	.92	.92	.93	.81	.83	.87	.92	.93	.92	.87	.87

Note. N = 97. Female is a dummy-coded variable: 0 = male, 1 = female. Correlations greater than |.20| are significant at $p < .05$. Correlations greater than |.29| are significant at $p < .01$.

79% C, 60% to 69% D, and below 60% F. Point biserial information for each test revealed no bad questions. Further information regarding the exams can be obtained from the authors upon request.

In addition to performance, we also collected a measure of students' interest in the course material using a 9-item scale (e.g., "I think the field of psychology is very interesting," "I really enjoy this class," "I plan on taking more courses in psychology," $\alpha = .92$ and .92). This measure of interest has been used in prior research (Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Hulleman et al., 2010), and is designed to capture students' emerging interest in psychology (Renninger & Hidi, 2011). Although this measure of interest was collected at the same time as the Time 3 measures of motivation, we used interest as an outcome in our analyses because, conceptually, interest is one of our key academic outcomes. Theoretical models of interest development (e.g., Hidi & Renninger, 2006) and empirical research (Harackiewicz et al., 2008) demonstrate that perceptions of competence and value are key antecedents of interest.

Procedure. The Time 1 survey was administered via an online survey during the second week of the semester. Students had a week to complete the survey, which included all Time 1 motivation items. During the 4th week of the semester, students took the first exam. During the 8th week of the semester, students completed the connection-frequency items (Time 2). During the 14th week of the semester, participants completed the Time 3 measures. In the 16th week, participants completed the final exam. Students earned course credit for completed surveys.

Results

Descriptive analyses. Our primary research question was whether the newly developed measure of connection frequency was related to self-reports of student motivation and course outcomes. Our hypothesis was that greater endorsement of the connection frequency items would be related to end-of-semester motivation, interest, and exam scores. Further, we examined whether connection frequency during midsemester provided an indirect pathway between initial motivation and final course outcomes.

The measure of connection frequency was normally distributed (skewness = −.11; kurtosis = .43), and the overall scale mean was just above the scale midpoint of 3.5 ($M = 3.9$ out of 6), which reflects making connections between 'sometimes' and 'often.' And 95% of the scores fell between 2.33 and 5.33 (see Figure S1). When inspecting the zero-order correlations (see Table 1), students' Time 2 self-reports of the frequency with which they connected the course material with their lives was moderately and positively correlated with Time 3 utility value ($r = .41$), expectancy ($r = .34$), and interest ($r = .46$), and negatively correlated with cost ($r = -.30$). Connection frequency was unrelated to final exam scores ($r = .05$). Thus, the measure of connection frequency is correlated in ways we would expect with other self-reported motivation variables, and had a sufficiently normal distribution and variance to warrant continuing to explore its role in the development of motivation during the semester.

Regression analyses. While the pattern of correlations provided preliminary evidence that connection frequency may contribute to learning outcomes, we ran a series of OLS regressions to investigate the unique pattern of relationships between Time 2 connection frequency and Time 3 outcomes, controlling for Time 1 covariates (measures of motivation, interest, exam 1 score, and gender). As shown in Table S1, connection frequency was a unique and significant predictor of utility value ($\beta = .25$, $sr^2 = .05$), cost ($\beta = -.21$, $sr^2 = .04$), and Time 3 interest ($\beta = .22$, $sr^2 = .04$). However, connection frequency was not a significant predictor of Time 3 expectancy or final exam score. We next tested whether Time 2 connection frequency served as an indirect pathway (ω) through which initial motivation could be related to course outcomes. Using the method outlined by Tofighi and MacKinnon (2011), we first regressed Time 2 connection frequency on the Time 1 covariates and found that Time 1 expectancy was the only significant predictor ($\beta = .24$, $sr^2 = .11$). Next, we utilized the prior regression equations to determine that Time 2 connection frequency served as an indirect pathway for Time 1 expectancy to contribute to changes in Time 3 utility value ($\omega = .08$, 95% CI [0.001, 0.20]), cost ($\omega = -.07$, 95% CI [−0.15, −0.01]), and Time 3 interest ($\omega = .08$, 95% CI [0.001, 0.19]). Finally, we examined whether connection frequency contributed to learning outcomes through its relationships with Time 3 motivation. This path model is displayed in Figure 1, and revealed that the relationship between connection frequency and interest could be explained by the increases in Time 3 utility value ($\omega = .15$, 95% CI [0.043, 0.287]).

Study 1 Discussion

Study 1 demonstrated initial validity evidence for our measure of the frequency with which students made connections between the course material and their lives. The measure was normally distributed, correlated with other self-reported motivation variables as expected, and contributed to students' motivation and learning outcomes. Consistent with this hypothesis, our measure of connection frequency was uniquely related to increases in expectancy, utility value, and interest, and decreases in cost, when controlling for prior measures of those variables. In addition, we found that connection frequency led to increased interest in psychology by increasing students' perceived utility value in the course. Although connection frequency did not operate as an indirect pathway between utility value and outcomes, as we had hypothesized, it did provide an indirect pathway between initial expectancy and outcomes. As a result, in Study 2 we examined whether the utility value intervention operates through success expectancies.

We also found in Study 1 that women outperformed men in the course. This replicates an emerging finding that women earn higher grades in school than men (e.g., Duckworth & Seligman, 2006; Voyer & Voyer, 2014). Our prior classroom-based intervention research demonstrates that the utility value intervention works better for low-performing students (Hulleman et al., 2010). Other interventions designed to promote perceptions of utility value for math and science found that the intervention differed depending upon students' gender and school performance (Rozek et al., 2015). These three findings set up the need to examine whether the intervention works differently for students based on demographics associated with being at-risk for poor performance (e.g., gender, initial performance levels), and experimental condition. We tested this possibility in Study 2.

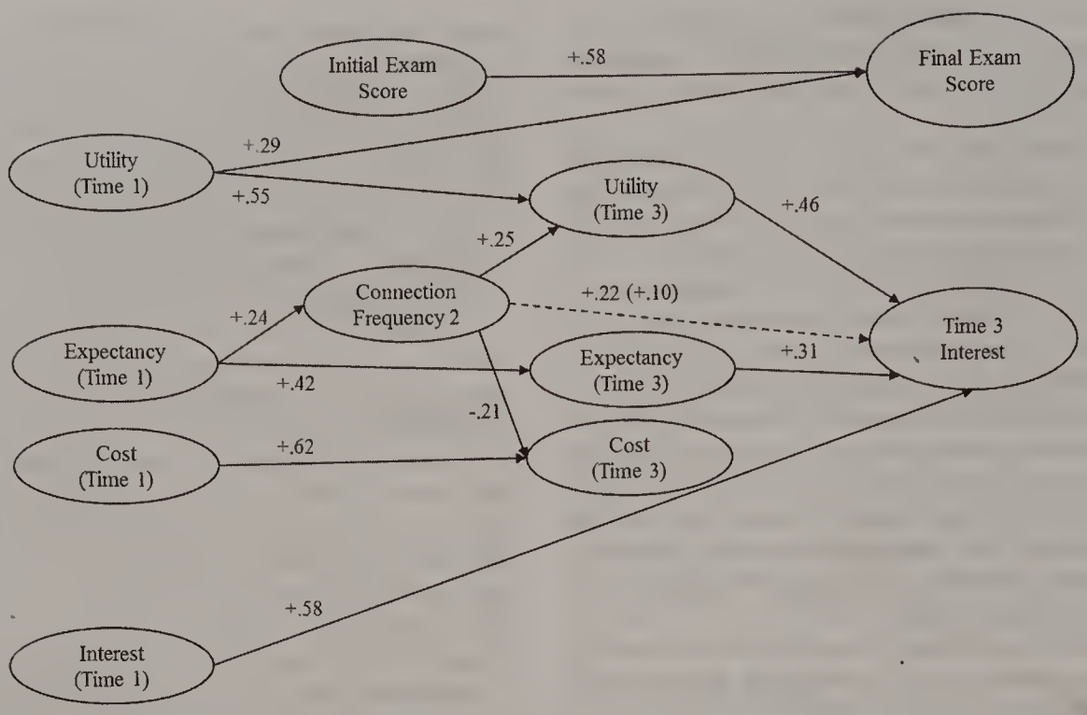


Figure 1. Path model of the relationships between connection frequency and learning outcomes in Study 1. $N = 97$. Values are standardized OLS regression coefficients that were statistically significant ($p < .05$). Only significant paths are shown. Regression equations also controlled for gender (see Table S1 and text for details). The direct effect of connection frequency on Time 3 Interest ($\beta = .22$, $p < .01$) was reduced to nonsignificant ($\beta = .10$, $p = .11$) when the Time 3 motivation measures were included in the model.

Study 2: A Longitudinal, Experimental Investigation

In Study 2, we conducted a double-blind, randomized classroom experiment that manipulated connection frequency by designing an enhanced utility value intervention that encouraged students to make more frequent connections between the course material and their lives. The goal of this enhanced utility value intervention was to increase the strength of the original utility value intervention by adding a new element. Although we hypothesized that connection frequency operated in the original utility value intervention, we wondered whether spontaneous connections may be uncommon or difficult to make (Bransford & Schwartz, 1999). We therefore utilized a related line of research on implementation intentions (e.g., Gollwitzer, 1999; Gollwitzer & Brandstatter, 1997), with the goal of increasing the connections students make between the material and their lives outside of the intervention.

When an individual forms an implementation intention, he or she specifies the when, where, and how an intended behavior will occur to promote goal attainment (Gollwitzer, 1999). The setting of these intentions provides a salient anchor for when a specific behavior should occur. In a study by Gollwitzer and Brandstatter (1997), participants were randomly assigned to either adopt implementation intentions for the completion of a self-reflection essay assigned over winter break, or were simply given the goal of turning in the essay. Students in the implementation intention condition were far more likely to complete the essay on time, and in less time, than a group of participants who were just given the goal to complete the essay.

We integrated the implementation intentions framework into the design of our enhanced utility value intervention. The enhanced condition included an opportunity for students to set implementation intentions to make connections between their lives and the course material on a routine basis during the semester (e.g., in class, while studying, when socializing). By adopting implementation intentions to make connections between the course material and their lives, we hypothesize that students will be more likely to actively seek connections in the specific situations that they identify. An increase in connections should promote deeper processing and engagement in learning, which in turn should enhance utility value, interest, and course performance. Because spontaneous connection-seeking does not always happen (Gentner et al., 2003), setting implementation intentions may nudge individuals toward this behavior.

Finally, we endeavored to make the interventions as easy to implement as possible. Both utility value interventions were designed so that they could be delivered via an online course management system used by the instructor and students. This allowed us, as researchers, to keep the instructor blind to students' experimental condition, and enabled students to participate in the intervention by using a familiar system. Although there was considerable set-up of the intervention required by the researchers, including randomizing of groups, the delivery of such an intervention to an entire class by a single instructor was done via an assignment through the online course management system. This approach is one solution for testing and scaling up psychological interventions in classrooms (Harackiewicz & Borman, 2014; Paunesku et al., 2015).

Method

Participants. Students in Study 2 were part of a separate subsample of students enrolled in the same two sections of intro-

ductory psychology used in Study 1. Of the original 589 students enrolled in the two sections, 388 were randomly selected to participate in Study 2. The final sample included 357 students who were over the age of 18, completed the final exam, and participated in the interventions. Similar to Study 1, the Study 2 sample was 70% female, 84% white (6% African American, 4% Asian), 84% nonpsychology majors, and 61% freshman (21% sophomore, 13% junior, 4% senior). The mean age of participants was 18.6 years.

Measures. All self-report and performance measures were identical to Study 1. Descriptive information on the scales, including reliabilities, can be found in Table 2.

Procedure. The procedures were nearly identical to Study 1, with two exceptions. First, the intervention prompts were delivered after the first and second exams. Second, instead of being measured at Time 2 (after the second exam), connection frequency was measured at the same time as the other motivation measures: Time 1 was the 2nd week of the semester and Time 3 was the 14th week. This allowed us to examine whether the intervention worked by increasing the frequency with which students connected the material to their lives. As in Study 1, students received course credit for completing the surveys and intervention prompts. Importantly, neither the instructor nor teaching assistants knew the specific content of the intervention, nor which students were assigned to which condition.

Intervention #1. Following the first exam, students were reminded in class to participate in the first intervention assignment and were given three days from the time that the links were available to complete the activity. Students then received web links to the intervention via email. The links were also posted in students' respective online group pages by the researchers. Upon clicking the web-link, participants were randomly assigned to one of three conditions: the control condition, the utility value condition, or the enhanced utility value condition. Thus, both the instructor and students were kept blinded to which students were in which conditions.

In the control condition (n = 119), participants received the following prompt: "Below is a list of the units covered in GPSYC 101 so far. For each topic, summarize what you know in about 1 or 2 sentences. We are not asking you to elaborate on the material, just to summarize the information that you can recall." Underneath the prompt were four text boxes labeled for each class unit (i.e., History, Careers, & Connections; Research; Biology & Behavior; and Memory).

Both the utility value (n = 116) and enhanced utility value (n = 122) conditions received the following prompt:

> In the space below, we would like you to write 1 to 2 paragraphs about how the material that you have been studying in GPSYC 101 relates to your life. We are not asking you to summarize the material, just to elaborate on its relevance to your life. So far, you have covered the following units in your class: History, Careers, & Connections; Research; Biology & Behavior; and Memory.

Below the prompt was a text-box for participants to type their short essays. This prompt was adapted from Hulleman and colleagues prior intervention prompts (Hulleman et al., 2010; Hulleman & Harackiewicz, 2009).

In addition, enhanced utility value participants were then sent to a new page featuring three additional prompts (see Appendix B in the Supplemental Online Material). The first prompt asked participants to identify the time and place where they might be able to think about the relevance of class material to their own lives. The second prompt asked participants to identify obstacles that might prevent finding the relevance of class material. The third prompt asked participants to identify strategies to overcome the obstacles identified in the second prompt.

Intervention #2. Following the 2nd exam, students in the control condition were given a pair of prompts in succession: "Choose one of the specific topics from above. In 1 to 2 paragraphs (75 to 125 words), summarize the details of the topic as best you can."

Table 2
Descriptive Statistics for Major Variables in Study 2

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Time 1 expectancy												
2. Time 1 utility value	.27											
3. Time 1 cost	−.38	−.36										
4. Time 1 connections	.30	.50	−.26									
5. Time 1 interest	.27	.79	−.38	.44								
6. Initial exam	.11	.01	−.05	.00	.03							
7. Time 3 expectancy	.40	.20	−.26	.16	.24	.34						
8. Time 3 utility value	.18	.67	−.33	.30	.65	.06	.43					
9. Time 3 cost	−.17	−.33	.54	−.15	−.42	−.27	−.42	−.44				
10. Time 3 connections	.19	.38	−.27	.47	.39	.10	.34	.51	−.34			
11. Time 3 interest	.07	.62	−.32	.30	.78	.10	.33	.75	−.50	.49		
12. Final exam	.09	.06	−.10	.00	.08	.63	.40	.11	−.31	.15	.13	
13. Female	.08	.33	−.19	.17	.33	.07	.12	.28	−.30	.21	.33	.12
Observed min.	4.75	2.17	1.67	1.00	1.44	35.00	1.50	1.00	1.83	1.50	1.67	39.00
Observed max.	8.00	8.00	7.00	6.00	8.00	76.00	8.00	8.00	8.00	6.00	6.00	100.00
Mean	6.46	6.12	3.52	3.77	6.11	61.62	6.34	6.47	5.97	3.99	3.99	79.48
SD	.81	1.07	.87	.84	1.14	7.15	1.01	1.08	1.13	.88	.88	9.40
α	.90	.92	.80	.88	.93	—	.93	.93	.86	.89	.93	—

Note. N = 357. Female is a dummy-coded variable: 0 = male, 1 = female. Correlations greater than |.10| are significant at $p < .05$. Correlations greater than |.15| are significant at $p < .01$.

Participants in the utility value and enhanced utility value conditions were given a pair of prompts about relevance: (a) "Choose a topic from above that is personally useful and meaningful to you. In 1 to 2 paragraphs (75 to 125 words), describe how learning about this topic is useful *to your life right now*" and, (b)

> Choose a topic from above that is personally useful and meaningful to you (it may be the same topic as before). In 1 to 2 paragraphs (75 to 125 words), describe how learning about this topic will be beneficial *to you in the future* (e.g., education, career, daily life).

Enhanced utility value participants were also given several items prompting reflection on their implementation intentions (see Appendix B in the Supplemental Online Materials). The reflection items asked students to recall what implementation intentions they had discussed in the previous intervention and to reflect on ways they could improve their strategies.

Results

Analytic plan. Our primary research question involved examining whether the intervention conditions would promote learning outcomes compared with the control condition (see Primary analyses). Our main hypothesis was that the utility value interventions would promote interest and achievement at the end of the semester compared with the control condition. We also hypothesized that the enhanced utility value condition would have an additional effect above and beyond the utility value condition. This first set of questions led us to conduct intent-to-treat, OLS regression analyses on the outcomes (interest, exam scores) as a function of the experimental conditions using hierarchical multiple regression. Second, we tested whether the intervention was more effective for students most at-risk for poor performance (see At-risk student analyses). This involved adding interaction terms between initial exam scores, gender, and the experimental conditions to the OLS regression models used in the primary analyses. Third, we tested whether the effects of the utility value interventions could be explained, at least partially, by increased motivation (i.e., expectancy, utility value, cost) and connection frequency. This would demonstrate that both motivation and connection frequency were indirect pathways through which the intervention impacted outcomes. This question was examined using path modeling and indirect effects analyses (see Indirect effects analyses). Finally, we conducted a fidelity analysis to examine whether students responded to the utility value writing prompts as expected (see Intervention fidelity analyses).

Descriptive analyses. A comparison of the unadjusted raw means in Table S2 reveals no significant difference on the Time 1 covariates (all Fs < 2.7, all ps $> .10$), indicating balanced randomization across the three conditions. Second, students successfully responded to the intervention prompts. There were no differences in the number of words written in the control ($M = 178.0$, $SD = 57.1$) or utility value conditions ($M = 182.8$, $SD = 57.7$; $d = .06$; $p = .57$), indicating students committed a similar level of effort and thinking in each condition. Example essays can be found in the Supplemental Online Materials (see Appendix C). In addition, the connection frequency variable was again normally distributed at both Time 1 (skewness = .19, kurtosis = .08) and Time 3 (skewness = .06, kurtosis = −.28), with a mean near the midpoint of the scale at both Time 1 ($M = 3.77$ of 6, $SD = .83$) and Time 3 ($M = 3.99$, $SD = .88$). However, as expected, there were small, raw mean differences in favor of the utility value intervention conditions compared with the control condition on Time 3 measures of interest ($d = .24$), expectancy ($d = .23$), utility value ($d = .24$), and final exam scores ($d = .23$), but not on cost or connection frequencies.

The pattern of correlations (see Table 2) was similar to Study 1 and revealed positive relationships between Time 1 connection frequency and Time 3 expectancy ($r = .16$) and utility value ($r = .30$), and a negative relationship with cost ($r = -.15$). Time 3 connection frequency was positively related to interest ($r = .49$) and final exam scores ($r = .15$).

Primary analyses. Did the utility value intervention conditions enhance academic outcomes compared with the control group? To answer this question, we used hierarchical, OLS regression to examine the effects of the interventions on interest in psychology and final exam scores. Regression allows us to examine unique relationships among the predictors of interest. We first examined intervention differences using an intent-to-treat model (Model 1) that included only the contrast codes for the experimental conditions. The utility value code compared whether both utility conditions were better than the control (control = −2, utility = +1, enhanced utility = +1), and the enhanced utility value code compared whether the enhanced utility value condition was better than the utility value condition (utility value = −1, enhanced utility value = +1). We did not include the covariates in Model 1 because there were no differences between experimental groups on the covariates, which meant that we could avoid the additional assumptions required when including covariates (Rosenbaum et al., 2002). However, adding the Time 1 motivation covariates (expectancy, value, cost) and connection frequency to the regression models predicting final exam scores and Time 3 interest did not alter the pattern of effects or statistical significance on either final exam scores or interest (see Model 3 in Tables S3 and S4).

Second, to examine whether the intervention was more effective for students at-risk of poor performance, we tested interactions between the intervention contrast codes and initial exam scores and gender in Model 2. Based on prior research (e.g., Hulleman & Harackiewicz, 2009; Hulleman et al., 2010) and Study 1, we identified male students who performed poorly on the first exam as most at-risk. Therefore, we added initial exam scores (mean-centered) as a continuous variable, gender (0 = male, 1 = female), four two-way interaction terms (utility contrast by gender, utility contrast by initial exams, enhanced contrast by gender, enhanced contrast by initial exams), and two three-way interaction terms to our model (utility contrast by gender by initial exams, enhanced contrast by gender by initial exams). This became Model 2.

Using the methods outlined by Aiken and West (1991), we probed significant interactions in three ways. First, we calculated predicted values at one standard deviation above and below the continuous moderator (in this case initial exam scores) by using the regression equation that contains the continuous variable. Second, we calculated simple slopes at one standard deviation above and below the continuous moderator to test for significant differences. Third, we calculated standardized mean differences between conditions at one standard deviation above and below the continuous predictor by estimating a regression equation with standardized predictors and standardized outcomes. Finally, to aid

interpretation of predicted values, we standardized the dependent variable when calculating predicted values so that they can be interpreted on a standardized metric.

Intent-to-treat analyses on learning outcomes. The analyses of Model 1 (see Tables S3 and S4) revealed a significant effect of the utility contrast on both final exam scores ($\beta = .12$, $sr^2 = .01$, $p = .03$) and interest ($\beta = .11$, $sr^2 = .01$, $p = .04$). Students randomly assigned to either utility condition performed better on the final exam ($M = 80.3$) and were more interested in psychology at the end of the course ($M = 6.14$) compared with students in the control condition ($M_{\text{Exam}} = 77.96$, $d = .25$; $M_{\text{Interest}} = 5.85$, $d = .24$). The difference between the utility and enhanced utility condition was not significant for either performance or interest.

At-risk student analyses. The analyses of Model 2 (see Tables S3 and S4) revealed a significant utility value contrast by initial exam interaction on both final exam scores ($\beta = -.22$, $sr^2 = .02$, $p = .01$) and interest ($\beta = 1.11$, $sr^2 = .01$, $p = .03$). As presented in Figure 2, when compared with students in the control condition, low-performing students in the utility value conditions performed better on the final exam ($d = .82$) and were more interested in the course material at the end of the semester ($d = .13$). Initial exam scores were also a significant predictor of final exam scores ($\beta = .60$, $sr^2 = .10$). Female students also reported more interest in psychology at the end of the semester than male students ($\beta = .32$, $sr^2 = .10$). Importantly, the enhanced utility value contrast was nonsignificant, indicating that there was no additional benefit of the enhanced utility value condition above and beyond the regular utility value condition (see Tables S3 and S4 for complete regression results).

The two-way interaction between the utility value contrast and initial performance on final exam performance was qualified by a significant three-way interaction between gender, initial performance, and the utility value contrast ($\beta = .19$, $sr^2 = .01$). As presented in Figure 3, the benefits of the utility value intervention appeared for low-performing male students who increased their exam performance by over three-quarters of a standard deviation in the utility value conditions compared with the control condition ($d = .76$). By the final exam, male students in the utility value conditions were performing as well as female students in the control conditions, which was equivalent to going from a C to a B

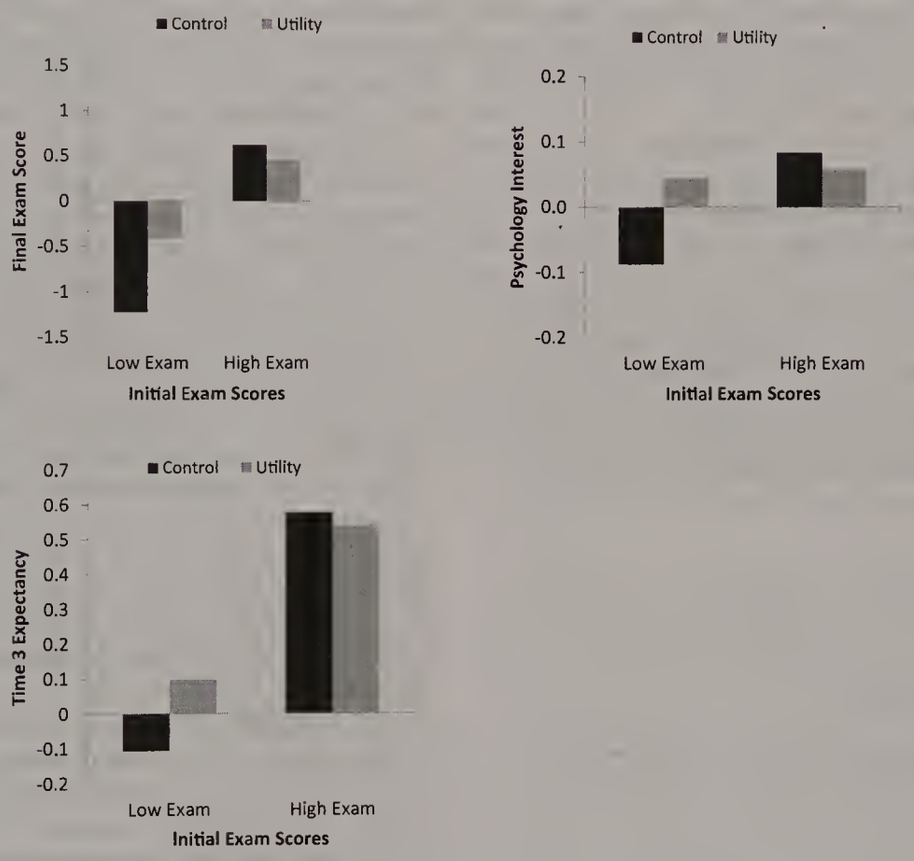


Figure 2. Interaction between the utility value interventions and initial exam scores on final exam scores, Time 3 interest in psychology, and Time 3 success expectancies in Study 2. $N = 357$. Predicted values for Low and High Exam were computed based on estimates for one standard deviation below (Low Exam) and above the mean (High Exam) on Initial Exam Scores (Aiken & West, 1991). We calculated standardized mean differences between the utility value and control conditions by using predicted values from a regression equation in which the outcome variables were standardized. Doing so results in predicted values that are in standardized units of both the predictor and the criterion (Cohen, Cohen, West, & Aiken, 2002). For example, in the upper left panel of this Figure, students with low expectancies in the control group had a predicted value of −1.22 on final exam scores, whereas low expectancy students in the combined utility value conditions had a predicted value of −0.41 on final exam scores. These values produce an adjusted, standardized mean difference of $d = 0.82$. See Model 2 in Tables S3 and S4 for complete details.

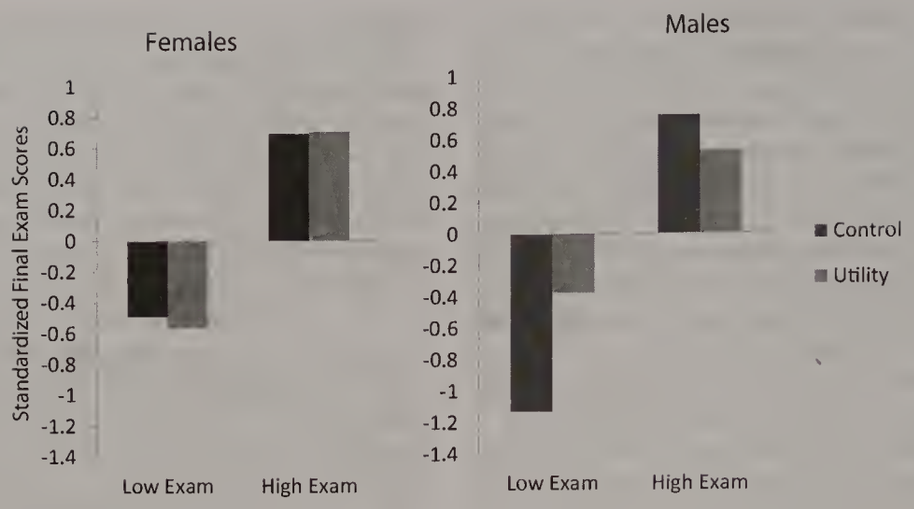


Figure 3. Three-way interaction between gender, initial exam scores, and utility conditions on final exam scores in Study 2. $N = 357$. Predicted values for Low and High Exam were computed based on estimates for one standard deviation below (Low Exam) and above the mean (High Exam) on First Exam scores (Aiken & West, 1991). See Model 2 in Tables S3 and S4 for complete details.

(see Figure 4). In other words, the students who traditionally perform most poorly in general education courses, males who initially struggle in the course, benefitted the most from the intervention. Importantly, the inclusion of the motivation covariates did not change the results (see Model 3 in Tables S3 and S4). The utility value conditions did not significantly affect high-performing students' exam scores, regardless of whether they were males or females, or low-performing females.

Indirect effects analyses. We also examined whether changes in connection frequency and expectancy-value-cost motivation were induced by the utility value interventions, and whether these changes contributed to further motivation and learning in the course. We also hypothesized that expectancy and perceptions of utility value might contribute to learning outcomes. We used path modeling within a multiple regression framework for two reasons. First, it matches the regression framework we used to analyze the intent-to-treat effects. Second, this was the same technique used in

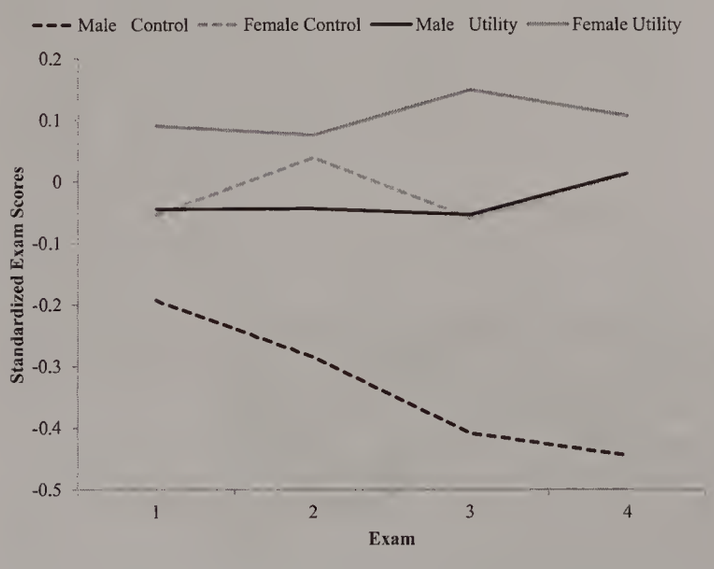


Figure 4. Unadjusted final exam scores by experimental condition and gender in Study 2. $N = 357$.

Study 1, with two exceptions. First, because we measured connection frequency at two time points, we first examined whether the interventions increased students' reports of connection frequency by regressing Time 3 connection frequency on the contrast codes, gender, initial exam scores, initial interest, and initial motivation (see Table S5 for regression results). Second, to test whether the motivation measures were pathways for the intervention effects, we included Time 3 measures of expectancy, value, cost, and connection frequency in the regression models predicting interest and final exam scores (see Model 4 in Tables S3 and S4).

Regressing Time 3 connection frequency on the Time 1 covariates and contrast codes for conditions revealed that initial interest in psychology ($\beta = .17$, $sr^2 = .01$), Time 1 connection frequency ($\beta = .35$, $sr^2 = .09$), and initial exam scores ($\beta = .10$, $sr^2 = .01$) were the only significant predictors. In terms of motivation, although the utility value interventions did not impact utility value or cost, they did impact students' success expectancies. When predicting Time 3 expectancies, there was a significant interaction between the utility value contrast and initial exam scores ($\beta = -.11$, $sr^2 = .01$). As presented in Figure 2, the utility value conditions increased low-performing students' expectancies compared with the control condition ($d = .20$), whereas there was no effect for high-performing students ($d = -.04$). Time 3 expectancy in turn was a significant predictor of final exam scores ($\beta = .21$, $p < .01$, $sr^2 = .02$). As presented in the top panel of Figure 5, low-performing male students in the utility conditions had higher exam scores than their counterparts in the control condition ($d = .76$), which was partially explained by an increase in expectancy for students who performed poorly on the first exam ($\omega = .15$, 95% CI [0.019, 0.322]). Importantly, these interaction effects were unaffected by the inclusion of expectancy in the regression model predicting final exam scores.

Although perceived utility value, cost, and connection frequency could not provide indirect pathways for the intervention effect (because they were not predicted by the intervention), the motivation variables were significant predictors of outcomes. Because we controlled for Time 1 measures of motivation and inter-

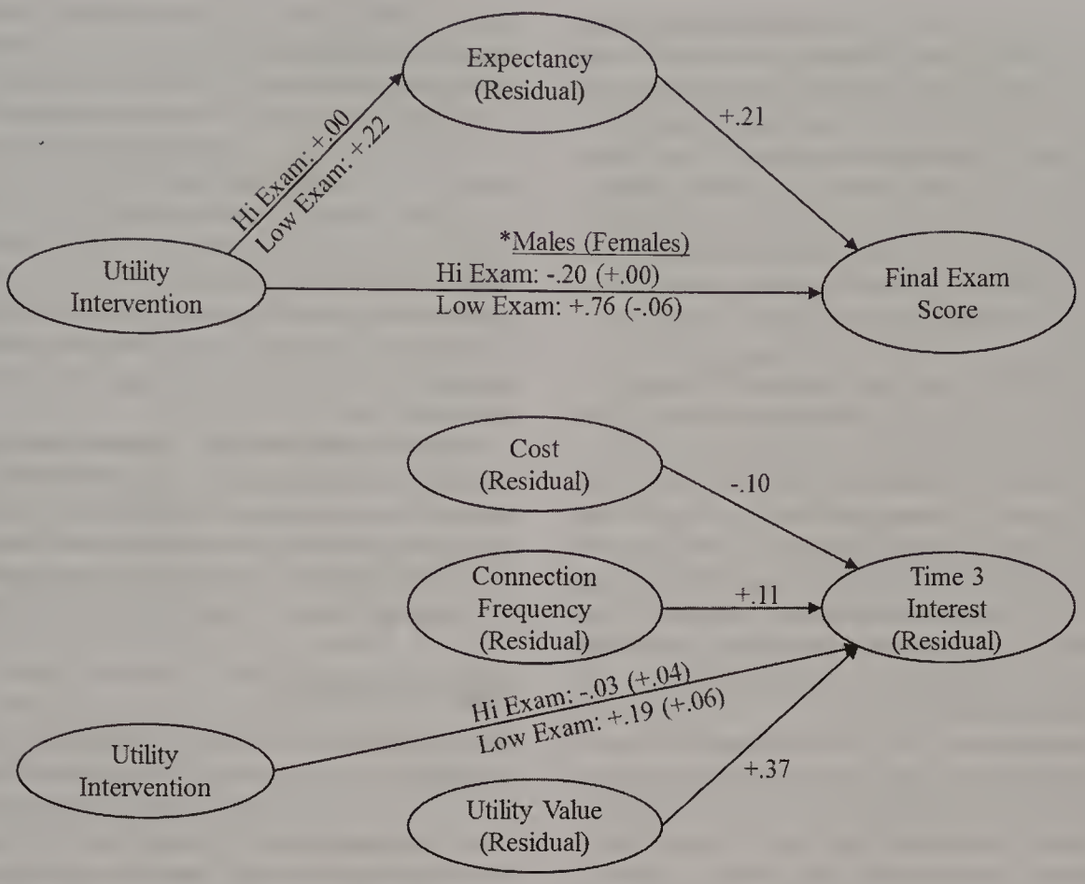


Figure 5. Path model of intervention effects on final exam scores (top) and interest (bottom) in Study 2. N = 357. Values are standardized OLS regression coefficients that were statistically significant ($p < .05$). Ovals with "(Residual)" in them are residual values having controlled for Time 1 measures. Other control variables included: Time 1 motivation (expectancy, utility value, cost, interest), gender, and initial exam scores. See text for details. The two-way interaction between the utility intervention and initial exam scores on interest was significant for students with low exam scores ($\beta = .19$), and reduced to nonsignificant ($\beta = .06$) when the motivation measures were included in the model. Predicted values for Low and High Exam were computed based on estimates for one standard deviation below (Low Exam) and above the mean (High Exam) on First Exam scores (Aiken & West, 1991). *The significant three-way interaction between the utility intervention, initial exam scores, and gender on final exam scores ($\beta = .19$, $sr^2 = .01$) revealed that low-performing males benefitted the most from the intervention ($d = .76$).

est, the Time 3 measures could be considered residual (or change) scores. As presented in the bottom panel of Figure 5, increases in perceived cost during the semester were associated with declines in interest during the semester ($\beta = -.10$, $sr^2 = .01$), whereas increases in perceptions of utility value ($\beta = .37$, $sr^2 = .05$) and connection frequency ($\beta = .11$, $sr^2 = .01$) were associated with increases in interest.

Intervention fidelity analyses. Although connection frequency was predictive of later interest, we did not increase connection frequency through our enhanced utility value manipulation. Because there were no differences between the utility and the enhanced utility conditions, this meant that the additional elements to the enhanced utility condition had no effect on outcomes above and beyond the utility writing. To further understand the effects of the intervention, we analyzed the extent to which participants responded to the intervention prompts as intended (i.e., intervention fidelity; see O'Donnell, 2008). To capture intervention fidelity, we identified three core components of the intervention prompts (Nelson et al., 2012): (a) the degree to which individuals provided satisfactory responses to their requisite prompts in all three conditions (indicating general compliance to the prompts across conditions); (b) the degree to which essays contained personalized connections between the material and their lives (indicating responsiveness to the utility value prompts); and (c) the degree to which individuals specified implementation intentions (indicating responsiveness to the enhanced utility value prompts). For the purposes of assessing intervention fidelity to the essay prompts, independent raters were trained on a brief rubric that contained three elements: writing quality (expected to be equal across conditions), personalization of connections (expected to be higher in the utility value conditions than in the control), and implementation intentions (expected to be higher in the enhanced utility condition compared with the control and utility condition).

Writing quality. To assess common elements of writing quality across all conditions, raters coded essays on a four-point rating scale that included 0 (*Less than a sentence*), 1 (*Typed incoherent thoughts or thoughts unrelated to the topic*), 2 (*A series of clear, unrelated sentences addressing the same topic*), 3 (*Groups of*

clear, related sentences addressing the same topic). Raters were allowed to use half points. Rater reliability was assessed using adjacent percent agreement, or the degree to which independent ratings were less than one point away on a rating scale (e.g., a rating of 2 and 2.5 would be considered agreement). In the case of disagreements, scores were averaged to compute the final rating. A score of 2 or higher on the writing quality scale was considered to be adequate fidelity. For writing quality, raters demonstrated 83% adjacent agreement across all interventions. The results indicated that the control group produced slightly lower quality essays during the first intervention ($M_{control}$ = 2.21, SD = 0.62; $M_{Utility}$ = 2.56, SD = 0.49; $M_{Enhanced}$ = 2.57, SD = 0.47), whereas the utility value condition was slightly lower during the second intervention ($M_{control}$ = 2.92, SD = 0.35; $M_{Utility}$ = 2.66, SD = 0.64; $M_{Enhanced}$ = 2.85, SD = 0.39). Despite the minor differences in writing quality, we note that all three groups received higher average ratings for the second set of essays, and all three group averages were between the highest points on the rating rubric. On average, 95% of essays were rated as having adequate fidelity, and only the control condition during first intervention (85%) was below 94%. These results suggested that students tended to write acceptable essays that addressed the prompted topics.

Personalized connections. To assess the degree to which personalization was present in both the utility and control essays, raters coded essays on a four-point rating scale that included 0 (*Essay is not focused on the self or a significant other*), 1 (*Essay implies or suggests personal importance, but does not say how*), 2 (*Essay references personal relevance rather than general*), 3 (*Essay references personal relevance and provides a strong example of why*). Essays were scored in the same manner as for writing quality. Raters demonstrated 92% adjacent agreement across all interventions. The results indicated that the control group displayed substantially lower personalization in their essays during the first intervention ($M_{control}$ = 0.04, SD = 0.34; $M_{Utility}$ = 2.24, SD = 0.67; $M_{Enhanced}$ = 2.21, SD = 0.69; d = 3.7 between the control and combined utility value conditions), as well as during the second intervention ($M_{control}$ = 0.02, SD = 0.15; $M_{Utility}$ = 2.34, SD = 0.74; $M_{Enhanced}$ = 2.50, SD = 0.55; d = 3.8 between the control and combined utility value conditions). On average, essays were rated as having adequate fidelity 89% of the time in the utility groups. The control essays were rated as being personalized 0.8% of the time, as would be expected. These differences suggest that the hypothesized driving feature of the intervention was present in the utility conditions, but not the control condition.

Implementation intentions. The enhanced utility value condition was created by including elements of implementation intention interventions. To that end, students were asked to respond to prompts about specific times to think about connections, about obstacles that might prevent them from thinking about connections, and about solutions to overcoming those obstacles. To measure fidelity of student responses in these cases, we used word counts for each individual prompt. Raters were asked to flag sentences that used language that could be used to answer any one of the three implementation intention prompts (i.e., times, obstacles, solutions; Rater Agreement = 100%). This approach was used because or initial analyses indicated that the presence (vs. absence) of implementation intentions was highly correlated with elaboration. Neither the control group nor the utility value group were coded as including implementation intentions (all mean word counts <1.55 words, median and mode word counts were 0 in both conditions at both time points). In contrast, the enhanced utility value prompts induced substantially more writing about implementation intentions during the first intervention (M_{Times} = *33.19*, SD = 25.27; $M_{Obstacles}$ = 31.26, SD = 23.61; $M_{Solutions}$ = 37.53, SD = 18.68), and slightly more during the second intervention (M_{Times} = 10.16, SD = 9.01; $M_{Obstacles}$ = 9.62, SD = 8.01; $M_{Solutions}$ = 10.37, SD = 7.97). To have adequate fidelity, students need to have indicated at least one time, one obstacle, and one solution. On average, control and utility essays were rated as having adequate fidelity 1% of the time, whereas essays in the enhanced utility condition were rated as having adequate fidelity 50% of the time.

Fidelity results summary. These results suggested that students generally demonstrated reasonable fidelity to the implementation intention prompts, although fidelity to the implementation intentions aspect of the enhanced utility condition was not quite as strong as fidelity to the personalization aspect of the regular utility condition. See Tables S6 and S7 for more details.

Study 2 Discussion

The results of Study 2 partially replicated prior research demonstrating that a theoretically guided intervention based on the expectancy-value framework could enhance student learning outcomes (Hulleman & Harackiewicz, 2009; Hulleman et al., 2010). As found in prior research, the utility value intervention worked best for students who were at-risk for poor overall course performance. On both interest in psychology and performance, the interaction between initial exam performance and the utility value intervention revealed positive effects for low performers and null effects for high performers. Furthermore, the three-way interaction on performance revealed that male students who had performed poorly on the first exam especially benefitted from the utility value intervention. We also replicated the effects of Study, which demonstrated that a new measure of connection frequency was a pathway through which students developed interest in psychology over the course of the semester. This directly supports our hypothesis that an important aspect of finding value in a topic, and eventually developing interest, is for students to make connections between the course content and their lives.

In addition to the utility value intervention used in prior research, we developed an additional intervention intended to increase connection frequency. The enhanced utility value intervention encouraged students to make more connections between the psychology they were learning and their lives during their daily routines. Unfortunately, our analyses did not reveal an additional benefit of the enhanced condition above and beyond the utility value condition. However, our investigation led to a surprising finding: instead of further bolstering students' utility value, as found in prior research (e.g., Hulleman et al., 2010), the utility value conditions increased students' expectancies for success. This was surprising because our theory and prior research had focused on perceptions of utility value as the primary mechanism of intervention effects. We consider these surprising findings in turn below.

What happened in the enhanced utility value condition? Our qualitative examination of the written responses to our intervention prompt indicated that students had, for the most part,

engaged in the intervention how we had intended. Students were prompted to make an implementation intention about when they would make connections, identify obstacles to making connections, and identify strategies to overcome those obstacles. These aspects of creating an intention to perform a specific behavior are aligned with the research literature in this area (e.g., Gollwitzer, 1999). So, what are we to conclude about the enhanced intervention? First, because this was our first attempt at creating this type of intervention, it is possible that our manipulation did not adequately activate behavioral commitment. The lack of effects on self-reported connection frequency seem to support this concern. One implication is to revise the intervention by including more aspects intended to activate behavioral commitment. However, instead of being an implementation issue, it is possible that our focus is on the wrong variable. Although our connection frequency measure may correlate with positive outcomes, it could be that when students are prompted to make connections on their own, that the quality of that connection also matters. Unfortunately, both our enhanced intervention prompts and connection frequency measures were solely focused on frequency and not quality. A second implication is to develop a different intervention that encourages students to make more high quality connections, similar to the types of connections that they are making when instructed to write about relevance, and to develop a measure that captures both quantity and quality of connections.

Why did the intervention boost success expectancies and not utility value? One obvious reason why we found effects on success expectancies not found in prior research is that success expectancies had not been previously hypothesized as a pathway of intervention effects, and thus had not been measured or analyzed in this way. However, mediation by success expectancies seems quite plausible. Several studies of the utility value intervention in both classroom-based field experiments and laboratories have found that the version of the utility value intervention used in this study, where students write about the connections they see as relevant to them, is most effective in boosting performance and interest for students with poor performance histories or low expectations (Durik et al., 2015). Thus, even though utility value was found to be a mediator in previous studies, it is quite plausible that the utility value intervention was working as a proxy for success expectancies. Conventional wisdom suggests that people like what they are good at and do better at what they like. This adage is supported by the finding that expectancies and values, when measured via self-reports, are positively correlated (Robbins et al., 2004). We found this to be true in both of our studies, with expectancy and utility value being moderately correlated (*r*s from .27 to .45). If this adage is true, it may be that the utility value interventions have been affecting expectancies and utility value, ultimately increasing both interest and performance. Yet prior investigations did not examine whether expectancies helped explain the intervention effects.

In addition to the surprising expectancy effect, it was also surprising that we did not find mediation of the intervention effect attributable to utility value in Study 2. Prior laboratory and field studies of the utility value intervention by Hulleman and colleagues (Hulleman et al., 2010; Hulleman & Harackiewicz, 2009) revealed that the effects of the intervention on outcomes could be explained, at least partially, by increases in perceptions of utility value. In our case there are at least three possibilities why our study might be different. First, this study was the first to manipulate utility value via an online assignment, so it may be that this difference changes how the intervention affects the mediating motivational mechanisms. In prior classroom studies, students either hand-wrote their responses to the intervention prompt in notebooks (2009), or had several weeks to hand-write or type a two to three page paper (2010, Study 2). In fact, there is some research to suggest that the physical act of writing activates different areas of the brain than typing the same text (Mueller & Oppenheimer, 2014). Further work is needed to examine this possible explanation.

Second, contextual differences in course instructor, instructional practices, or student characteristics could be responsible. The small sample of contexts (this is only the third published test of this intervention in the field) makes the influence of such contextual differences difficult to discern. However, there are important contextual aspects of this study worth noting. The instructor teaching the course is known for being highly motivating, and has received national teaching awards from The Princeton Review (2012) and the American Psychological Association (2012). Perhaps even more importantly for these particular findings, the student body at his university voted him as the best professor during 2013 (Jacobs, 2013). Evidence of his teaching prowess is also apparent in our survey data. Students reported making more frequent connections to psychology at the end of the semester than at the beginning. In fact, the change in connection frequency from the fourth to the thirteenth week was over a quarter of a standard deviation in the control conditions (d = .25) and combined utility value conditions (d = 0.27). As a point of reference, this change was much larger than for either expectancy or value, which both slightly decreased over the semester (both ds = −.13). These contextual differences could change the way the intervention affects motivational dynamics. For example, with such a strong base for value and making connections being provided by the context, the measure of utility value may have reached a ceiling which could not be adjusted by the intervention. In this strong value context, the utility value intervention may have emboldened students to believe they could succeed because they trusted the teacher to make the content not only interesting, but also learnable. As demonstrated in K through 12 settings, trust in school is an important predictor of increased achievement scores (Bill & Melinda Gates Foundation, 2010; Bryk & Schneider, 2002), and future research could investigate trust in relation to the utility value intervention.

Third, it can be difficult to capture mediation for many reasons. There are examples of other interventions which are based on social psychological theory (e.g., growth mindsets, belonging uncertainty, values affirmation) that have effects on outcomes but do not always capture effects on measured variables. Self-reports are prone to biases, such as social desirability and reference bias, which are sensitive to context (Duckworth & Yeager, 2015). This is particularly true for utility value as students' responses are likely affected by the specific unit or topic that students are studying, which varies from week to week.

General Discussion

The results of Study 2 add to the growing body of literature that social psychological interventions in general can promote student

learning outcomes, and that utility value interventions in particular can be beneficial. In addition to replicating prior research on utility value, we extended this growing body of work in several ways. To further understand the mechanism of the utility value intervention effect, we both measured (Studies 1 and 2) and manipulated (Study 2) one process of the utility intervention effect. This method can provide additional validity evidence in support of the mediating mechanism (Baron & Kenny, 1986; Sigall & Mills, 1998). Specifically, we hypothesized that the frequency with which students made connections between the material and their lives throughout the semester might be one way that the utility value intervention increases motivation and performance. By continually making connections, students might be energizing their study behavior and integrating their knowledge in deeper ways (Bransford & Schwartz, 1999). Although we were unable to successfully manipulate this mechanism in Study 2, we found that making more connections between the material and students' lives was positively related to expecting to doing well in the course, and valuing the course material. In addition, students who made more connections perceived fewer costs for learning the material. Because this was an exploratory investigation of this measure, future research is needed to further validate the measure and understand its relationship with motivational processes and learning outcomes.

Implications for Theory and Research

At a theoretical level, the current research lends additional support for understanding motivation and achievement in educational contexts using an expectancy-value framework. In particular, the role that expectancy and utility value both play in determining key academic outcomes within the context of interventions was further elucidated. Although prior theoretical work allowed for expectancies and values to be positively related, it was only through experimental research that we learned that a utility value intervention can actually increase expectations for success. We also uncovered an important proximal process, or mechanism, through which the utility value intervention has its effects: connection frequency. In both studies, students who reported more frequently seeing connections between the course material and their lives reported more interest in the material at the end of the semester. Further, in Study 1, this link between connection frequency and outcomes was explained by a concomitant increase in perceptions of utility value. Although this finding is correlational, it corroborates our inclination that the frequency of connections is an important aspect of finding value, and developing interest, in academic content. As in past research, encouraging students to make a connection at a single point in time through a utility value intervention boosted utility value and learning outcomes. In extending prior work, we demonstrated that making frequent connections between the material and students' lives also boosts utility value and learning outcomes.

Our findings are also consistent with models of interest development (Hidi & Renninger, 2006; Renninger & Hidi, 2011), which posit that perceiving value in a particular domain or activity is a crucial aspect of developing an enduring interest. In addition to providing support for the role of value in interest development through an intervention study, we also found a new pathway for interest development: connection frequency. Although not explicitly outlined in their four-phase model of interest, connection frequency is likely related to two important factors in this model: perceived knowledge and value. By making more connections, students are building additional knowledge about how learning content relates to their lives, and also creating a foundation for perceiving personal value in a topic. Future work will need to further develop our understanding of connection frequency and quality in interest development. Importantly, this link to interest development, and in particular understanding which factors are amenable to manipulation within the classroom context, are especially important for educational practice.

Implications for Practice

On the surface, our effort to further explore the mechanisms of the utility value intervention is a theoretical question. Why should practitioners care why an intervention works so long as it works? However, our work uncovered a theoretically surprising finding related to this question: The utility value intervention, in this context, increased low performing students' outcomes by virtue of enhancing success expectancies. One interpretation of this finding is that an intervention designed to enhance value actually enhances students' expectancies. For practitioners, solving their local challenges involves aligning the sources of the problem (i.e., students with low expectancies) with an intervention that targets that source. This means looking beyond the surface features of the intervention to understand the mechanics of how the intervention works (Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012).

In addition, this research is an example of designing an intervention in an online environment to facilitate achieving both research and practice goals. On the research end, the online delivery enabled us to randomize students behind the scenes so that both students and the instructor were blind to differences in intervention activities across students. Data entry and cleaning were minimized, and the data were immediately available to us. We were able to inform the instructor which students had completed the interventions and surveys so that he could assign course credit, and students had a common means of accessing the 'assignment' regardless of condition. On the practice end, the online environment minimized the impact on instructional time (students completed the intervention and surveys on their own time outside of class). An online environment is also a cost-effective means of scaling psychological interventions, as most undergraduate institutions use some version of course management software within which the intervention and surveys can be imbedded.

At-Risk Students and Academic Achievement

Students disengage from school, and eventually drop out, as a result of a number of factors, including lack of academic preparation, poor knowledge of effective learning strategies, and low motivation (e.g., Allensworth & Easton, 2007). Poor performance in introductory and general education courses, especially those taken early in students' academic career (e.g., middle school prealgebra or college-level general education courses), can have an especially salient impact on their academic trajectories (e.g., Casillas et al., 2012). Interestingly, one group that has consistently underperformed in school has been males. A recent meta-analyses demonstrated that females outperform males in school, and this performance gap cuts across grade-levels, subject areas, and pub-

lication year (Voyer & Voyer, 2014), indicating that this is not a recent phenomenon. A separate review indicated that gender differences in intelligence, personality, and motivation partially explained this performance gap (Spinath, Eckert, & Steinmayr, 2014). In particular, girls are more self-disciplined than boys, which leads to increased learning and academic achievement (Duckworth & Seligman, 2006).

The results from Study 2 are consistent with this emerging work, and provide additional evidence that the utility value intervention helps students at risk for underperformance. In both of our studies, male students performed more poorly than their female counterparts. In Study 2, the utility value intervention reduced this gap by over 75%. These results also align with other social psychological interventions that boost academic achievement of at-risk student groups (e.g., Aronson, Fried, & Good, 2002; Cohen, Garcia, Apfel, & Master, 2006; Walton & Cohen, 2011). For example, the utility value intervention has boosted the performance of students who initially doubted their ability to succeed in high school science (Hulleman & Harackiewicz, 2009) and undergraduate science (Hulleman, An, Hendricks, & Harackiewicz, 2007), first-generation under-represented minority students in college biology (Harackiewicz et al., 2015), and underperforming undergraduate psychology students (Hulleman et al., 2010). Further research will need to determine whether the gender effect was simply a proxy for identifying low-performing students, or whether it identified an important difference between male and female students in particular.

Limitations and Future Directions

There are some important limitations to this study, including those that are common to field experiments, such as studying intervention effects within a single context, which necessarily constrains generalizations about effectiveness (cf. Shadish, Cook, & Campbell, 2002; Shavelson, Phillips, Towne, & Feuer, 2003). In our two studies, four limitations stand out as particularly important for future research. First, our measure of connection frequency, which was newly developed for this study and central to our research questions about the pathways of the utility value intervention effects, needs further validation. How does it correlate with other measures of motivation, and how sensitive is it to differences in teaching style or learning content? Furthermore, the measure does not capture the quality of connections that students make. It is highly likely that students who make high-quality connections will benefit more than students who make low-quality connections to their lives. There are at least two possible reasons for this. From a motivation perspective, making higher quality connections could deepen students' desire to digest the material and engage in learning. From a neuroscience perspective, research shows that new experiences can be associated with existing memories when these experiences are strongly activated (e.g., McGaugh, 2000). When two neural pathways are activated in tandem, the intensity in activation can trigger a process known as long term potentiation (Purves et al., 2001). Long-term potentiation leads to new synaptic connections between the two pathways, resulting in the two pathways being activated together in future experiences of either. This would mean that making a connection between course content and a common daily experience (e.g., working at a job) could then lead to the activation of that content whenever the daily experience reoccurs. Regardless of the underlying reason, the importance of quantity and quality of connections could be examined in future research. For example, student essays in response to the utility value intervention prompts could be coded for quality of connections, and then linked to outcomes (e.g., Hulleman & Cordray, 2009). Alternatively, an accompanying student self-report measure of connection quality could be assessed so that the independent effects of connection quantity and quality could be examined.

The final three limitations are related to the possible reasons why the intervention effect sizes in our experimental study were not enhanced as we had hoped. Second, because of the potential influence of classroom contextual factors (e.g., instructor, instructional practices, peer norms), future research needs to be conducted within a wide variety of classrooms and instructional styles. Ideally, intervention effects would be examined in enough classrooms to examine between classroom differences in the effectiveness of the utility value intervention. Not every intervention will replicate in every context, whether due to implementation challenges or other issues (e.g., Dee, 2015). Thus, the need for independent replication work is necessary for the utility intervention just as it is for any other intervention in education contexts.

Third, this was the first time we used an online medium to deliver the intervention, and we adapted the intervention so as to reduce the time burden on students. As a result, intervention dosage, and quite possibly intervention strength, was reduced compared with prior studies of the utility value intervention. In this study, students were asked to write less text and to do it less frequently than prior versions of the utility value intervention. For example, in the previous college psychology study published by Hulleman et al. (2010), students completed two take-home essays of 1 to 2 pages, compared with two online essays of 2 to 3 paragraphs. Thus, both of these factors (highly motivating classroom context and intervention strength) may have conspired to mute the salience of the intervention on students' perceptions of utility value. However, despite these implementation changes, the overall effect sizes on learning outcomes in this study (*d*s from .23 to .24) were similar to other value interventions (Lazowski & Hulleman, 2015). Future research could systematically vary these implementation factors to identify necessary levels of dosage and frequency required to obtain utility value intervention effects.

Fourth, the control condition used in Study 2, as well as in the other published randomized field experiments of the utility value intervention (Harackiewicz et al., 2015; Hulleman & Harackiewicz, 2009; Hulleman et al., 2010), was not a do-nothing control group. Rather, the control condition consisted of asking students to summarize the material they had been learning recently in the course. In the research literature on the cognitive psychology of learning, this control condition is known as summarization and has been found to enhance learning (Dunlosky et al., 2013). Essentially, these studies (including ours) tested a motivation intervention in comparison to a cognitive intervention, and found that the motivation intervention produced better outcomes. This means that the effect size for the utility value intervention, both in the research presented here as well as prior published work, likely has been underestimated because the comparison group contained a cognitive intervention. To obtain a more pure effect size, future research could use a more inert comparison group.

Conclusion

The methods in this study demonstrate the value of experimental tests of psychological theories. Without such intervention studies, we would know very little about what happens in classrooms when we try to enhance student motivation (cf. Shavelson et al., 2003). Our combined longitudinal and experimental approach provides initial validity evidence for the role of connection frequency and motivation in explaining utility value intervention effects. More generally, this research contributes additional validity evidence to the growing body of research related to the impact of social-psychological interventions on educational outcomes (Lazowski & Hulleman, 2015; Yeager & Walton, 2011). When such theoretically guided interventions are thoughtfully implemented within academic contexts, surprisingly strong and consistent effects have been found on interest in academic topics, course performance, and persistence (Wilson, 2006). Importantly, this research demonstrates that these effects are not "magic," but rather rely on targeted psychological mechanisms that can gain influence over time (Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009; Garcia & Cohen, 2011). The hopeful message is that, by engineering the psychological situation, educational practitioners can significantly impact student learning and development.

References

Acee, T. W., & Weinstein, C. E. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *Journal of Experimental Education, 78,* 487–512. http://dx.doi.org/10.1080/00220970903352753

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on track and graduating in Chicago Public High Schools*. Chicago, IL: Consortium on Chicago School Research. Retrieved December 17, 2007.

American Psychological Association. (2012). *Psychology's top honors: Div.* 2 (Society for Teaching of Psychology). Robert S. Daniel Teaching Excellence Award (four-year college): David B. Daniel, PhD. Retrieved from http://www.apa.org/monitor/2012/09/top-honors.aspx

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84,* 261–271. http://dx.doi.org/10.1037/0022-0663.84.3.261

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African-American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38,* 113–125. http://dx.doi.org/10.1006/jesp.2001.1491

Ash, K. (2008). Promises of money meant to heighten student motivation. *Education Week*. Retrieved on February, *14,* 2008.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64,* 359–372. http://dx.doi.org/10.1037/h0043445

Barney, S. T. (2007). Capitalizing on the self-references effect in general psychology: A preliminary study. *Journal of Constructivist Psychology, 20,* 87–97. http://dx.doi.org/10.1080/10720530600992915

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173

Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology, 80,* 706–722. http://dx.doi.org/10.1037/0022-3514.80.5.706

Barron, K. E., & Hulleman, C. S. (2015). Expectancy-value-cost model of motivation. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 8, pp. 503–509). Oxford, UK: Elsevier Ltd. http://dx.doi.org/10.1016/B978-0-08-097086-8.26099-6

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*. Research paper retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Boekaerts, M. (2002). *Motivation to learn*. Educational Practice Series #10. International Bureau of Education. Bellegarde, France: SADAG. Retrieved from http://www.ibe.unesco.org/fileadmin/user_upload/archive/publications/EducationalPracticesSeriesPdf/prac10e.pdf

Bong, M. (2001). Role of self-efficacy and task-value in predicting college students' course performance and future enrollment intentions. *Contemporary Educational Psychology, 26,* 553–570. http://dx.doi.org/10.1006/ceps.2000.1048

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24,* 61–100.

Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist, 34,* 75–85. http://dx.doi.org/10.1207/s15326985ep3402_1

Brown, E. R., Smith, J. L., Thoman, D. B., Allen, J. M., & Muragishi, G. (2015). From bench to bedside: A communal utility value intervention to enhance students' biomedical science motivation. *Journal of Educational Psychology, 107,* 1116–1135. http://dx.doi.org/10.1037/edu0000033

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage Foundation.

Burke, L. A., & Hutchins, H. M. (2007). Training transfer: An integrative literature review. *Human Resource Development Review, 6,* 263–296. http://dx.doi.org/10.1177/1534484307303035

Casillas, A., Robbins, S., Allen, J., Kuo, Y. L., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology, 104,* 407–420. http://dx.doi.org/10.1037/a0027180

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313,* 1307–1310. http://dx.doi.org/10.1126/science.1128317

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324,* 400–403. http://dx.doi.org/10.1126/science.1170769

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied multiple regression/correlation analyses for the behavioral sciences* (3rd ed.). London, UK: Routledge.

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology, 104,* 32–47. http://dx.doi.org/10.1037/a0026042

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4899-2271-7

Dee, T. S. (2015). Social identity and achievement gaps: Evidence from an affirmation intervention. *Journal of Research on Educational Effectiveness, 8,* 149–168. http://dx.doi.org/10.1080/19345747.2014.906009

Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98,* 198–208. http://dx.doi.org/10.1037/0022-0663.98.1.198

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational pur-

poses. *Educational Researcher, 44,* 237–251. http://dx.doi.org/10.3102/0013189X15584327

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. http://dx.doi.org/10.1177/1529100612453266

Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology, 99,* 597–610. http://dx.doi.org/10.1037/0022-0663.99.3.597

Durik, A. M., Hulleman, C. S., & Harackiewicz, J. M. (2015). One size fits some: Instructional enhancements to promote interest don't work the same for everyone. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 49–62). Washington, DC: American Educational Research Association.

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology, 98,* 382–393. http://dx.doi.org/10.1037/0022-0663.98.2.382

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 74–146). San Francisco, CA: Freeman.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin, 21,* 215–225.

Flake, J., Barron, K. E., Hulleman, C. S., McCoach, D. B., & Welsh, M. E. (2015). Understanding cost: The Forgotten component of expectancy-value theory. *Contemporary Educational Psychology, 41,* 232–244.

Garcia, J., & Cohen, G. L. (2011). A social psychological approach to educational intervention. In E. Shafir (Ed.), *Behavioral foundations of policy*. Princeton, NJ: Princeton University Press.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95,* 393–405.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355. http://dx.doi.org/10.1016/0010-0285(80)90013-4

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54,* 493–503. http://dx.doi.org/10.1037/0003-066X.54.7.493

Gollwitzer, P. M., & Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology, 73,* 186–199. http://dx.doi.org/10.1037/0022-3514.73.1.186

Grays, M. P. (2013). *Measuring motivation for coursework across the academic career: A Longitudinal invariance study.* [Unpublished doctoral dissertation.] James Madison University, Harrisonburg, VA.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94,* 562–575. http://dx.doi.org/10.1037/0022-0663.94.3.562

Harackiewicz, J. M., & Borman, G. (2014, April). Scaling up social psychological interventions to address achievement gaps in education. Symposium at the annual conference of the American Educational Research Association. San Francisco, CA.

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen, C. J., Blair, S. S., Rouse, D. I., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology, 106,* 375–389. http://dx.doi.org/10.1037/a0034679

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2015). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspp0000075

Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100,* 105–122. http://dx.doi.org/10.1037/0022-0663.100.1.105

Harackiewicz, J. M., Rozek, C. R., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science, 23,* 899–906. http://dx.doi.org/10.1177/0956797611435530

Harackiewicz, J. M., Tibbetts, Y., Canning, E., & Hyde, J. S. (2014). Harnessing values to promote motivation in education. In S. A. Karabenick & T. C. Urdan (Eds.), *Advances in motivation and achievement* (Vol. 18, pp. 71–105). Bingley, UK: Emerald Publishing. http://dx.doi.org/10.1108/S0749-742320140000018002

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70,* 151–179. http://dx.doi.org/10.3102/00346543070002151

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41,* 111–127. http://dx.doi.org/10.1207/s15326985ep4102_4

Hulleman, C. S., An, B., Hendricks, B., & Harackiewicz, J. M. (2007, June). *Interest development, achievement, and continuing motivation: The pivotal role of utility value.* Poster presented at the Institute of Education Sciences Research Conference, Washington, DC.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The Role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2,* 88–110. http://dx.doi.org/10.1080/19345740802539325

Hulleman, C. S., Durik, A. M., Schweigert, S., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology, 100,* 398–416. http://dx.doi.org/10.1037/0022-0663.100.2.398

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102,* 880–895. http://dx.doi.org/10.1037/a0019506

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326,* 1410–1412. http://dx.doi.org/10.1126/science.1177067

Jacobs, P. (2013, October 1). America's hottest professor is more than just a pretty face—Here's why students are crazy about his class. *Business Insider.* Retrieved from http://www.businessinsider.com/interview-with-americas-hottest-professor-david-daniel-2013-9

Johnson, M. L., & Sinatra, G. M. (2013). Use of task-value instructional inductions for facilitating engagement and conceptual change. *Contemporary Educational Psychology, 38,* 51–63. http://dx.doi.org/10.1016/j.cedpsych.2012.09.003

Kohn, A. (1999). *Punished by rewards.* Boston, Massachusetts: Houghton Mifflin.

Kosovich, J. J., & Hulleman, C. S. (2016). A utility value framework: Task-goal relevance in achievement motivation. Manuscript under review.

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2015). A practical measure of student motivation: Establishing validity evidence for the expectancy-value-cost scale in middle school. *The Journal of Early Adolescence, 35,* 790–816.

Lazowski, R. A., & Hulleman, C. S. (2015). Motivation interventions in education: A meta-analytic review. *Review of Educational Research, 86,* 602–640.

McGaugh, J. L. (2000). Memory—A century of consolidation. *Science, 287,* 248–251. http://dx.doi.org/10.1126/science.287.5451.248

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25,* 1159–1168. http://dx.doi.org/10.1177/0956797614524581

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research, 39,* 374–396. http://dx.doi.org/10.1007/s11414-012-9295-x

Newby, T. J. (1991). Classroom motivation: Strategies of first-year teachers. *Journal of Educational Psychology, 83,* 195–200. http://dx.doi.org/10.1037/0022-0663.83.2.195

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78,* 33–84.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science, 26,* 784–793. http://dx.doi.org/10.1177/0956797615571017

Peters, D. L., & Daly, S. R. (2013). Returning to graduate school: Expectations of success, values of the degree, and managing the costs. *The Journal of Engineering Education, 102,* 244–268. http://dx.doi.org/10.1002/jee.20012

Pike, A. G., & Dunne, M. (2011). Student reflections on choosing to study science post-16. *Cultural Studies of Science Education, 6,* 485–500. http://dx.doi.org/10.1007/s11422-010-9273-7

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A. S., McNamara, J. O., . . . Williams, S. M. (2001). Plasticity of mature synapses and circuits. In D. Purves, G. J. Augustine, D. Fitzpatrick (Eds.), *Neuroscience* (2nd ed.). Sunderland, MA: Sinauer Associates. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK10878/

Renninger, K. A., & Hidi, S. (2011). Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist, 46,* 168–184.

Renninger, K., & Wozniak, R. H. (1985). Effect of interest on attentional shift, recognition, and recall in young children. *Developmental Psychology, 21,* 624–632. http://dx.doi.org/10.1037/0012-1649.21.4.624

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138,* 353–387. http://dx.doi.org/10.1037/a0026838

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130,* 261–288. http://dx.doi.org/10.1037/0033-2909.130.2.261

Rosenbaum, P. R., Angrist, J., Imbens, G., Hill, J., Robins, J. M., & Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science, 17,* 286–327. http://dx.doi.org/10.1214/ss/1042727942

Rozek, C. S., Hyde, J. S., Svoboda, R. C., Hulleman, C. S., & Harackiewicz, J. M. (2015). Gender differences in the effects of a utility-value intervention to help parents motivate adolescents in mathematics and science. *Journal of Educational Psychology, 107,* 195–206. http://dx.doi.org/10.1037/a0036981

Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13,* 23–52. http://dx.doi.org/10.1023/A:1009004801455

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher, 32,* 25–28. http://dx.doi.org/10.3102/0013189X032001025

Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review, 2,* 218–226. http://dx.doi.org/10.1207/s15327957pspr0203_5

Simons, J., Vansteenkiste, M., Lens, W., & Lacante, M. (2004). Placing motivation and future time perspective theory in a temporal perspective. *Educational Psychology Review, 16,* 121–139. http://dx.doi.org/10.1023/B:EDPR.0000026609.94841.2f

Spinath, B., Eckert, C., & Steinmayr, R. (2014). Gender differences in school success: What are the roles of students' intelligence, personality, and motivation? *Educational Research, 56,* 230–243. http://dx.doi.org/10.1080/00131881.2014.898917

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121,* 371–394. http://dx.doi.org/10.1037/0033-2909.121.3.371

The Princeton Review. (2012). *The best 300 professors.* Farmingham, MA: Author.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8,* 247–261. http://dx.doi.org/10.1037/h0074898

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods, 43,* 692–700. http://dx.doi.org/10.3758/s13428-011-0076-x

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140,* 1174–1204. http://dx.doi.org/10.1037/a0036620

Vroom, V. H. (1964). *Work and motivation.* New York, NY: Wiley.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331,* 1447–1451. http://dx.doi.org/10.1126/science.1198364

Wigfield, A., & Cambria, J. (2010). Expectancy-value theory: Retrospective and prospective. In T. C. Urdan & S. A. Karabenick (Eds.), *The decade ahead: Theoretical perspectives on motivation and achievement (Advances in motivation and achievement*; Vol. 16, pp. 74–146). Bingley, UK: Emerald Group Publishing Limited.

Wilson, T. D. (2006). Behavior: The power of social psychological interventions. *Science, 313,* 1251–1252. http://dx.doi.org/10.1126/science.1133017

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81,* 267–301. http://dx.doi.org/10.3102/0034654311405999

Received January 8, 2015
Revision received June 16, 2016
Accepted June 17, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 3, 405–424

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000141

New Evidence on Self-Affirmation Effects and Theorized Sources of Heterogeneity From Large-Scale Replications

Paul Hanselman
University of California, Irvine

Christopher S. Rozek
University of Chicago

Jeffrey Grigg
Johns Hopkins University

Geoffrey D. Borman
University of Wisconsin–Madison

Brief, targeted self-affirmation writing exercises have recently been offered as a way to reduce racial achievement gaps, but evidence about their effects in educational settings is mixed, leaving ambiguity about the likely benefits of these strategies if implemented broadly. A key limitation in interpreting these mixed results is that they come from studies conducted by different research teams with different procedures in different settings; it is therefore impossible to isolate whether different effects are the result of theorized heterogeneity, unidentified moderators, or idiosyncratic features of the different studies. We addressed this limitation by conducting a well-powered replication of self-affirmation in a setting where a previous large-scale field experiment demonstrated significant positive impacts, using the same procedures. We found no evidence of effects in this replication study and estimates were precise enough to reject benefits larger than an effect size of 0.10. These null effects were significantly different from persistent benefits in the prior study in the same setting, and extensive testing revealed that currently theorized moderators of self-affirmation effects could not explain the difference. These results highlight the potential fragility of self-affirmation in educational settings when implemented widely and the need for new theory, measures, and evidence about the necessary conditions for self-affirmation success.

Keywords: values affirmation, replication, stereotype threat, achievement gap, middle school

Supplemental materials: http://dx.doi.org/10.1037/edu0000141.supp

One potentially promising approach to reducing persistent racial/ethnic achievement gaps is to tackle their social–psychological dimensions, including the negative consequences of stereotype threat and other identity threats in school. Because identity threats have detrimental consequences for marginalized groups in many academic settings (Steele, Spencer, & Aronson, 2002), such approaches can have substantial impacts. For instance, brief reflective writing exercises conducted in school settings can provide large and lasting benefits for theoretically threatened groups, such as African American and Hispanic middle-school students (Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009; Sherman et al., 2013), women in a college physics course (Miyake et al., 2010), and first-generation college students (Harackiewicz et al., 2014).

How robust are these effects? Although benefits of seemingly simple interventions suggest great potential, researchers caution that these techniques are "not magic" (Yeager & Walton, 2011). By their nature, the interventions target specific interactions between individuals and their social context and, therefore, critical differences in intervention delivery, individual students, or social

This article was published Online First August 8, 2016.

Paul Hanselman, School of Education, University of California, Irvine; Christopher S. Rozek, Department of Psychology, University of Chicago; Jeffrey Grigg, School of Education, Johns Hopkins University; Geoffrey D. Borman, Departments of Educational Leadership and Policy Analysis and Sociology, University of Wisconsin–Madison.

Research reported in this article was supported by the U.S. Department of Education through Grant R305A110136 to the University of Wisconsin–Madison (Principal Investigator [PI]: Geoffrey Borman) and Grant R305B120013 to the University of California, Irvine (PI: Greg Duncan); the Spencer Foundation Grant 201500044 (PI: Geoffrey Borman); and the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number P01HD065704 (PI: Greg Duncan). The content is solely the responsibility of the authors and does not necessarily represent the official views of the supporting agencies.

We are grateful to Geoffrey Cohen for sharing implementation materials and to Geoffrey Cohen, Gregory Walton, Joshua Aronson, John Protzko, and Valerie Purdie-Vaughns for advice during the design of this project. We appreciate helpful comments on previous versions of this article from Greg Duncan and Judy Harackiewicz, as well as seminar participants at the advisory board meeting of the Irvine Network on Interventions in Development (January 2015) and the Irvine Motivation Meeting (January 2015). Jaymes Pyne provided specific research assistance related to this article.

Correspondence concerning this article should be addressed to Paul Hanselman, School of Education, University of California, Irvine, 3200 Education Building, Irvine, CA 92697-5500. E-mail: paul.hanselman@uci.edu

contexts may lead to substantial variability in effectiveness. As a result, one must gauge the impact of these interventions in diverse settings and, to the extent that there are meaningful differences in effects, assess whether theorized moderators explain these differences. If heterogeneous effects follow theoretically predictable patterns, then these interventions have a clear role in improving educational outcomes and reducing achievement gaps. However, if heterogeneity remains unpredictable, then the immediate value of these interventions is less clear.

Theorized heterogeneity also complicates the fundamental enterprise of independent replication, which is increasingly recognized as necessary to build firm scientific understanding in psychology as in other fields (Ioannidis, 2005, 2012; Pashler & Harris, 2012). If the impacts of social–psychological interventions depend on seemingly subtle differences in delivery, individuals, and social contexts, then discrepant replication results may reflect predictable differences in effectiveness across diverse settings. On the other hand, mixed results may be due to unpredictable study-specific differences, such as unrecognized moderators or sampling variation. This distinction is especially difficult to disentangle when studies are conducted by different investigators and with different populations in different contexts. As a result, initial replication efforts of affirmation interventions in educational settings—which demonstrate success (e.g., Sherman et al., 2013), challenges (e.g., Kost-Smith et al., 2012), and failure (e.g., Dee, 2015)—raise questions about both the size and variability of these effects when implemented broadly. In particular, do theorized moderators explain differences in self-affirmation benefits? This study provides unique evidence on this question by reporting on a new large-scale test of self-affirmation effects and comparing these results to a previous effort in the same setting.

Self-Affirmation: Theory and Promise

This study is informed by theories of social identity threats, which create particular challenges for members of marginalized social groups in school (Steele et al., 2002). For instance, Black and Hispanic students are subject to *stereotype threat* in academic settings, in which they face the threat of conforming to or being judged by negative stereotypes about their racial/ethnic group (Steele & Aronson, 1995). The experience of stereotype and other identity threats leads to poorer academic performance through a variety of psychological responses, including stress, anxiety, and vigilance (Schmader, Johns, & Forbes, 2008), and may contribute to longer term disengagement and a "downward spiral" of performance (Cohen & Garcia, 2008). Because these stereotype threats uniquely apply to groups subject to negative academic stereotypes, they may account for portions of the widening of racial achievement gaps in school.

Stereotype threats are pernicious because students are affected by virtue of membership in a marginalized group (regardless of whether or not they endorse a negative stereotype, as long as they are aware of it), and broad social stereotypes are difficult to change. Instead, the goal of many social–psychological interventions is to reduce the harm that existing threats cause by shifting how students view themselves and/or their social world (Wilson, 2011). The example we consider is a set of brief writing exercises that ask students to reflect on meaningful personal values, such as family, friends, music, or sports. Following their initial presentation (e.g., Cohen, Garcia, Apfel, & Master, 2006; Cohen et al., 2009; Sherman et al., 2009), we refer to these activities as *self-affirmation* interventions throughout this article, reflecting the goal to allow students to "reaffirm their self-integrity" (Cohen et al., 2006, p. 1307). Similar interventions have also been described as "values affirmation" (e.g., Cook, Purdie-Vaughns, Garcia, & Cohen, 2012; Harackiewicz et al., 2014; Shnabel, Purdie-Vaughns, Cook, Garcia, & Cohen, 2013).

Self-affirmation interventions are believed to restore an individual's sense of worth in the face of threats related to social identity, thus mitigating detrimental stress responses (Steele, 1988). Because individual identities are complex, individuals "can maintain an overall self-perception of worth and integrity by affirming some other aspect of the self, unrelated to their group" (Sherman & Cohen, 2006, p. 206). Threats to academic identity experienced by minority members in school can be muted by focusing attention on other specific aspects of identity (Critcher & Dunning, 2015; Sherman & Cohen, 2006; Steele, 1988; Walton, Paunesku, & Dweck, 2012). Reflection on important values provides a psychological buffer against the full brunt of detrimental stereotype threats in school, and because of the potentially recursive nature of threat and poor performance, subtle buffering early on may lead to substantial benefits over time (Cohen & Garcia, 2008; Cohen et al., 2009; Taylor & Walton, 2011; Walton, 2014).

Geoffrey Cohen and his colleagues have developed these theoretical ideas alongside specific classroom writing activities to promote self-affirmation via reflection on important values. Each activity takes 15–20 min and is conducted by classroom teachers several times during the school year; the timing emphasizes critical moments such as the beginning of the school year and potentially stressful evaluative milestones. Consistent with theoretical expectations, these activities did not significantly impact White students' academic performance, who likely experienced relatively little academic identity threat (Walton & Cohen, 2003). However, the effects on grade point average (GPA) for 7th grade African American and Hispanic students were substantial and persistent (Cohen et al., 2006; Cohen et al., 2009; Cook et al., 2012; Sherman et al., 2013). Remarkably, the benefits of the intervention reduced the racial achievement gap in the targeted course by 40% (Cohen et al., 2006, p. 1307), which suggests great potential for this approach to address educational disparities that are associated with identity threat processes.

What mediates these effects? Critcher and Dunning (2015) presented recent laboratory evidence for an "affirmation as perspective" model, in which self-affirmations "expand the contents of the working concept—thus narrowing the scope of any threat" (p. 4). Working concept refers to the salient identities that make up one's self-concept in consciousness at any point in time. When aspects of identity are threatened, working self-concept tends to constrict, amplifying the negative experiences of that threat. However, if a broader working concept is maintained, then threats associated with a specific aspect of identity are less salient. It stands to reason that self-affirmation in school expands the contents of self-concept for students subject to academic stereotypes, thus reducing attention to the threat and muting the stress responses that lead to poorer performance.

Empirical tests of mediators in middle school settings have been mixed. Cook et al. (2012) reported impacts of self-affirmation on Black students' level and variability of sense of belonging in

school, which indicate effects on students' construal of their social environments, but the authors argued that these effects are "not a mechanism in the sense of mediation" (p. 483). Similarly, Sherman et al. (2013) reported impacts on higher levels of construal and a more robust sense of social belonging, whereas Cohen et al. (2006) reported decreases on a measure of cognitive activation of racial stereotype, yet neither found evidence that these effects mediated the impact of self-affirmation. Shnabel et al. (2013) found that writing about social belonging mediated some of the self-affirmation benefits; however, Tibbetts et al. (2016) did not replicate this result in another setting and instead found that writing about independence mediated some of the affirmation benefits.

The self-affirmation writing exercises have been implemented in at least four middle school field settings beyond the original one. Figure 1 summarizes both the positive impacts from early field trials within three schools (Cohen et al., 2006; Sherman et al., 2013) and smaller and sometimes nonstatistically significant estimates in large-scale, multischool replications (Borman, Grigg, & Hanselman, 2016; Dee, 2015).[1] The latter are well-powered studies conducted by independent research teams, and their results raise questions about the fundamental sources of variability in self-affirmation effects. Unfortunately, many features of the research settings varied in these studies and little implementation information is available to isolate the impact of specific differences. For instance, the study conducted by Dee (2015) illustrates

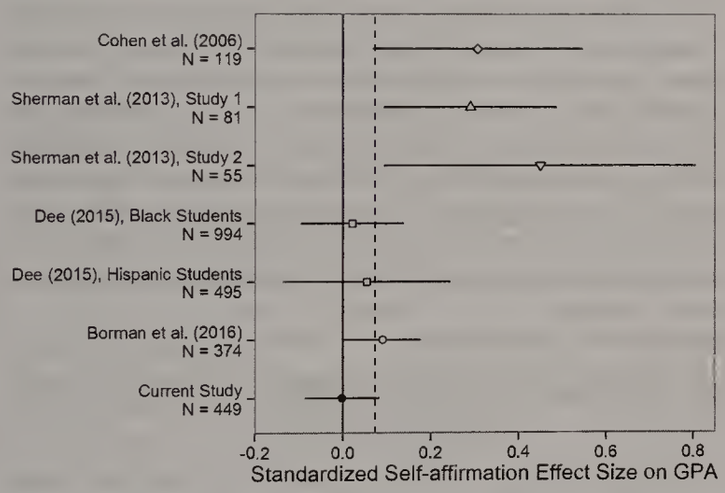


Figure 1. Estimated effects of self-affirmation writing exercises on middle school grade point average (GPA). Source: authors' calculations; see Table A1 in the online supplemental materials for specific references. Symbols plot reported effect sizes for potentially stereotyped groups (African American and/or Hispanic students) for the first year of the self-affirmation intervention, and lines represent 95% confidence intervals (+/−1.96 standard errors). Shapes represent distinct school or district contexts. For instance, Sherman et al. (2013) Studies 1 and 2 were conducted in different schools in different states. Dee (2015) reports subgroup results from the same sample of Philadelphia-area schools. The dashed line represents the overall mean effect size (0.07), calculated by weighting individual estimates according to the inverse of their squared standard error. The impact estimates are lower in the large-scale replication studies (Borman et al., 2016; Dee, 2015, and Current Study), but these differences could reflect heterogeneous effects across local context, research team, and implementation. This article investigates two effects observed within the trial conducted in a single school district (represented by circles), for which context and procedures were consistent.

multiple potentially relevant changes across research efforts. For one, it was conducted in schools with substantial minority student populations; these are contexts where self-affirmation may be less effective (Hanselman, Bruch, Gamoran, & Borman, 2014). For another, it recruited an unusually representative sample of students (a 94% consent rate), which could account for dampened impacts if the students not typically included in other studies benefit less from the intervention. These preliminary results suggest the need for more precise consideration of where, for whom, and under what conditions self-affirmation is beneficial.

Theoretical Moderators of Self-Affirmation Effects

Psychological theory posits that self-affirmation is beneficial in specific circumstances (Cohen & Sherman, 2014; Yeager & Walton, 2011), highlighting the need to identify the necessary and sufficient "preconditions" for its benefits in educational settings (Cohen et al., 2006). Null results emphasize this point, because existing theory provides post hoc explanations but not clear insight into when, where, and why self-affirmation might not have worked (e.g., see Harackiewicz, Canning, Tibbetts, Priniski, & Hyde, 2015). And of course if moderators were well understood, then studies would likely not have been fielded in such unsuccessful contexts.

In surveying potential self-affirmation moderators, the literature points to three relevant domains: features of the delivery of the activities, individual characteristics of the participating students, and aspects of the social context. First, specific features of the delivery of the brief self-affirmation intervention are hypothesized to be necessary for students to benefit. For example, Critcher, Dunning, and Armor (2010) found that self-affirmation exercises were only effective when introduced before a threat or before participants became defensive in response to a threat, which suggests that it is important to implement self-affirmation exercises before stressful events in school in order to short-circuit negative recursive cycles (see also Cohen & Garcia, 2014; Cook et al., 2012). Qualities of presentation that shape how students perceive the writing activities—such as making participants aware that exercises are beneficial (Sherman et al., 2009) or externally imposing affirmation (Silverman, Logel, & Cohen, 2013)—may mute self-affirmation benefits. Conversely, researchers have argued that the activity is most beneficial when presented as a normal classroom activity (Cohen & Sherman, 2014; Purdie-Vaughns et al.,

[1] The summary presented in Figure 1 should be viewed as an informal account of previous self-affirmation impacts in middle school settings. A formal and more expansive meta-analysis will certainly be useful in the future as more independent evidence emerges, but our specific purpose in collecting these estimates was to provide context for the current study. We therefore focus only on studies in middle schools that report self-affirmation effects on overall GPA relative to an alternate activity. These criteria rule out studies at other levels (e.g., Miyake et al., 2010), those that consider other outcomes (e.g., Cook et al., 2012; Study 1), and those without a nonself-affirmation control group (e.g., Cook et al., 2012; Study 2). Similarly, we omit the study by Bowen, Wegmann, and Webber (2013) because reported values do not include an overall estimate of impacts on GPA (that study reports offsetting impacts on initial GPA and slope over time; inspection of their Table 3 and Figure 1 suggests this study would contribute a small negative effect on overall GPA to our summary if included). We include detailed information about the source of represented estimates in Appendix Table A1 in the online supplemental materials.

2009) and when promoting specific types of writing (e.g., Shnabel et al., 2013). Finally, the type of control group used has also been suggested as an implementation-based moderator of the effects of self-affirmation. The typical control group, which asks students to write about nonimportant values, has the potential to undermine students' confidence if they write about activities in which they have low ability whereas other control writing prompts, which are more neutral or open-ended, might allow control participants to spontaneously affirm themselves (McQueen & Klein, 2006).

Second, numerous individual difference variables have been hypothesized to make students more vulnerable to stereotype threat and thus moderate the effects of self-affirmation, including identifying with a negatively stereotyped group, being knowledgeable about self-relevant negative stereotypes, and caring about doing well in school (Aronson et al., 1999; Cohen & Sherman, 2014; Shapiro & Neuberg, 2007). Therefore, although all negatively stereotyped minority students might be helped by self-affirmation, subgroups that are even more highly negatively stereotyped, such as Black males (Eagly & Kite, 1987; Purdie-Vaughns & Eibach, 2008; Sidanius & Pratto, 1999) or the lowest-achieving minority students (Cohen et al., 2009), might benefit most from self-affirmation.

Finally, context variables are hypothesized to moderate self-affirmation benefits. Social characteristics, such as group composition and environmental cues, influence the behavior and performance of stereotyped students (Dasgupta, Scircle, & Hunsinger, 2015; Inzlicht & Ben-Zeev, 2000; Murphy, Steele, & Gross, 2007). The effectiveness of self-affirmation approaches depends on the identity threats "in the air" in a particular setting (Steele, 1997), and the hypothesized recursive benefits are theorized to depend on relatively rich learning environments for threatened students to take advantage of as they are buffered from perceived threats (Cohen & Sherman, 2014). Because self-affirmation is theorized to disrupt stereotype threat processes, settings in which threats are more likely to be experienced may provide the greatest opportunity for benefits. For instance, although self-affirmation reduced gender disparities in performance in an introductory college physics course (Miyake et al., 2010), it was not beneficial in introductory science settings in which gender gaps and stereotype threat were not present (Lauer et al., 2013). Theory and empirical evidence also suggest that minority students attending schools in which their group is poorly represented and in which there are large racial achievement gaps benefit most from self-affirmation (Cohen & Garcia, 2014; Hanselman et al., 2014).

In summary, psychological theory posits moderators of self-affirmation effects in several domains, but evidence for specific moderators is limited because the data to test these theories are lacking, especially in applied educational settings. This means that mixed evidence of self-affirmation benefits may be due to theorized variation in how the activities were delivered, individual characteristics, or social contexts. In particular, very little is known about how to translate theorized constructs and laboratory manipulations into measures of the relevant moderating features as they occur in applied settings. Moreover, it is impossible to isolate specific relevant differences between the independent field trials to date, which have been conducted in different contexts with different populations and different procedures. Nonetheless, interrogating potential moderators is key to assessing both the underlying theory of self-affirmation and its likely practical impact. To the extent that a priori hypotheses predict heterogeneity, these results would confirm theory and point to where these strategies have the most potential to improve student outcomes. On the other hand, it is possible that mixed self-affirmation results are not explained by currently theorized moderators, which would imply the need for greater and more specific inquiry into the necessary conditions for success.

A New Self-Affirmation Replication Study

Given variable evidence of impacts in applied settings, we tested the effects of brief, in-class self-affirmation writing exercises for 7th grade students on subsequent academic outcomes in a new double-blind randomized experiment in a sample of more than 1,200 students in one Midwestern school district. We sought to learn whether similar benefits could be attained in a different setting, both in terms of geographic location and scale of implementation.

The Original Study

The original self-affirmation study in a middle school setting was first reported by Cohen et al. (2006), with supplemental analyses elsewhere (Cohen et al., 2009; Cook et al., 2012; Shnabel et al., 2013). We replicated the procedures in the original experiments as described below. Cohen and his colleagues originally reported several substantively important features of self-affirmation intervention on student outcomes: substantial persistent benefits for "negatively stereotyped" students (African American and Hispanic students) on GPA; significantly higher benefits for low-performing African American students; an improved trend in grades throughout the year; and no benefits for European American students. Our primary focus was on the first finding, representing the highly policy-relevant main impact of the intervention on negatively stereotyped groups. The impact for African American students ranged from 0.21 to 0.34 GPA points across individual experiments and across courses (Cohen et al., 2006, p. 1308).

The Previous Independent Replication in the Current Research Setting

The immediate precedent for the current self-affirmation replication is the study reported by Borman et al. (2016). That study was the first successful independent replication of the benefits of self-affirmation benefits in middle schools. The researchers reported statistically significant benefits for "potentially threatened" students (Black and Hispanic) on 7th grade GPA across all schools in the district. Like the original study, term-specific GPA data revealed a less negative trend for potentially threatened students in the self-affirmation condition, and no benefits for "not potentially threatened" students (White and Asian). Some results deviated from the original patterns. For one, the impacts were smaller, with an impact of 0.065 cumulative GPA points; the confidence interval for this estimate was [0.001, 0.128], which excludes all impact estimates from the original study. The authors speculated that this difference may have been at least partially related to the challenges of implementing at scale. Also, the replication found no evidence of an interaction between the intervention and prior achievement. In supplemental analyses, researchers reported that the treatment

benefits in this scale-up were concentrated in a subset of schools hypothesized to have the most threatening environments for potentially threatened groups, based on the numerical presence and relative academic standing of these students (Hanselman et al., 2014).

The Current Study

The current study was designed to replicate both the original self-affirmation study (Cohen et al., 2006) and the previous successful independent replication (Borman et al., 2016). Three key features of this design provide unique insights into the effects of self-affirmation in educational settings. First, procedures followed those in the original study, including intervention materials, as we detail below. The study therefore is an example of a well-powered "close" replication of the effects of self-affirmation for potentially threatened groups in middle school (Brandt et al., 2014). Moreover, given the scale of the research, the study contributes important evidence about the general promise of these interventions to improve minority students' achievement.

The second key feature of the study is that it was conducted in the same setting as a previous randomized trial of self-affirmation, in the same district and schools, by the same research team, with the same research protocols. In the current study, we ask whether these middle school scale-up results were replicated, and we use comparisons across studies to test theorized sources of heterogeneity. Because features of the study corresponded closely to those in the previous one (Borman et al., 2016; see Table A2 in the online supplemental materials for a summary), the *within-setting* comparisons across the two studies allow for much more specific tests of moderation than comparisons between settings. A recent precedent for such a within-setting comparison is provided by Harackiewicz et al. (2015), who found different affirmation effects in a college setting and discussed several potential explanations for the difference. We exploit a similar pattern to conduct comprehensive tests of theorized sources of heterogeneity.

A third contribution of this study is that we collected information on self-affirmation implementation, including qualitative features of students' responses to the exercises. These data provide an unprecedented picture of the experience of the self-affirmation activities when they are implemented in an entire school district. And, in combination with information about individual student characteristics and features of the social context, this information supports unique tests of the theorized sources of heterogeneity.

Building on the unique empirical features of this research, we addressed three sequential research questions. Our first question was, What was the effect of the self-affirmation intervention in the new large-scale implementation? Because we found no evidence of benefits, we asked our second question: Were estimated effects substantively and significantly different from the impacts for the students from a previous study in the same setting? Given meaningful and detectable differences, our third question was, Why was the same intervention seemingly beneficial for targeted students in one implementation but less so in the next?

The third research question is the most theoretically important, but it also is the most challenging. To preview our approach, we drew on the theory underlying the design of the interventions to conduct a series of tests of potential explanations for differences in effects across studies. Based on hypothesized moderators of the impacts of self-affirmation, these explanations fall into three broad classes: characteristics of implementation, individuals, and social context. We then conducted a series of empirical tests of these potential explanations to assess which, if any, explained the differences in experimental impact estimates.

Method

The Large-Scale Self-Affirmation Studies

All data were generated or collected as part of two randomized trials of self-affirmation writing activities among 7th grade students. The research was conducted through a partnership with the school district, which recognized large racial achievement gaps and was interested in strategies to improve the performance of minority students. District administrators provided support to the project, and principals at all 11 regular middle schools agreed to participate. Given this support, study implementation involved researchers (who provided training and activity materials), school learning coordinators (who coordinated the site-specific logistics, including scheduling), and teachers (who implemented the activities in their classrooms). The involvement of educators in diverse roles approximated how the exercises would be likely to be implemented if adopted as a universal district initiative.

Throughout this article we refer to the first study, conducted with 7th grade students in 2011–2012, as "Cohort 1" and the second study, conducted in 2012–2013, as "Cohort 2." The focus of this article is on the new evidence on self-affirmation effects provided by Cohort 2; no results from this study have been reported previously. In order to compare results across the two studies, we also conducted new analyses of participants in Cohort 1, including documenting impacts in 8th grade. We therefore detail aspects of both the new study (Cohort 2) and the previous one (Cohort 1).

The general outline of both studies was similar, as follows: Research activities began in the summer with parallel contact at each of district's 11 middle schools. After confirming authorization from the principal and identifying an appropriate setting for the writing exercises with each school's learning coordinator, research staff provided a training session for the 7th grade instructional teams at each school. During the 30-min training session, a member of the research staff introduced the study as research about 7th grade students' experiences, beliefs, and social-emotional learning. The researcher described the mechanics of implementation and reviewed the teacher implementation script. Teachers administered the writing exercises during normal class time with materials provided by the research team and the completed exercises were returned to the research team for recording. After the school year, the district provided administrative data, including transcript and demographic information. No study activities were conducted after the 7th grade year, but additional administrative data on 8th grade performance were collected after the following year.

Below we highlight the core features of the intervention, with a focus on similarities and differences between the two studies. Appendix Table A2 in the online supplemental materials provides a summary.

Self-Affirmation Intervention and Implementation

The self-affirmation intervention procedure followed Cohen et al. (2006). Seventh grade students completed a short (15–20 min) writing prompt as part of normal class activities several times during the school year. We identified four time points for the self-affirmation writing interventions. These provided a consistent template for the district, but scheduling varied according the formative assessment dates in individual schools. The time points were (a) at the start of the school year, in the week prior to formative fall standardized assessments; (b) in November, in the week prior to the state's standardized achievement test for accountability purposes; (c) in the winter, in the week prior to a midyear language skills formative assessment; and (d) in the spring, in the week prior to the final formative assessment of the year. Based on the evidence that self-affirmation exercises are most effective earliest in the school year (Cook et al., 2012), we provided school officials with the option of omitting the winter exercise to reduce logistical challenges; four schools did so for Cohort 1 and two did so for Cohort 2.

The activities were administered by teachers in the classroom using scripts provided by the original research team. Forty-five teachers were involved in Cohort 1, 44 were involved in Cohort 2, and 33 were consistent across both studies; teacher changes reflected exits from the school, reassignments, and looping (teachers moving grades along with students). The intervention activities were completed in a classroom setting determined by the school's learning coordinator to be the most appropriate for the writing exercises: in language arts classes at seven schools and homeroom period at four (constant across both cohorts). Homeroom periods were abbreviated classes with nonacademic curricula, including activities related to socioemotional standards. In either case, exercises were implemented among all 7th graders in these regular classrooms by their classroom teachers.

The activities were packets of 3–4 pages with prompts and spaces for individual writing responses. They were identical on the cover sheet, which included the student's name. On subsequent pages the exercises varied by randomly assigned condition (for consented students; all nonconsented students, including newly enrolled students without a personalized packet, completed the procedural/neutral control prompts). The treatment condition, following the original study, prompted students to reflect on values (such as friends, family, music, or sports) that were important to them. The precise format of the treatment exercise varied throughout the year to avoid repetition. There were two randomly assigned control conditions: one focused on values, in which students are asked to select least important values from the same list presented to treatment students and explain why they may be important to someone else, and a second devoted to various procedural writing prompts, such as describing summer activities or explaining how to open a locker (we refer to these prompts as "neutral," as they do not explicitly concern values). The latter control branch was introduced after the first administration in the Cohort 1 study, so all control students in the first cohort received the "Least Important Values" prompt for the first exercise. Because we found no evidence of differences between control conditions in either cohort nor evidence that these differences explain differential impacts, we combined both control groups in our main analyses.

Individualized packets were prepared for every student in the district based on classroom rosters and distributed to teachers ahead of implementation. The priority in implementation procedures was to promote an environment in which students engaged in the genuine self-reflection about aspects of identity that is hypothesized to lead to self-affirmation benefits. One implication, following previous research, is that activities were to be conducted as a normal part of classroom activity; this point was stressed in the teacher training and implementation scripts. However, the fact that teachers implemented the activities independently in their own classrooms created challenges for documenting precise features of implementation, as we discuss below.

We also instructed teachers to avoid representing the activities as evaluative, to avoid reference to external research, and to avoid presenting the activities as beneficial. These guidelines were based on theory and empirical evidence (Cohen & Sherman, 2014; Silverman et al., 2013), with the caveat that there is little existing guidance about how these features translate into best practice for teachers in established educational settings. For instance, anecdotal feedback from teachers highlighted some tension between these theoretical ideals and integration into classroom activities. For many students and some teachers, the medium of the activities—a personalized packet completed individually—led to a default perception of the activities as a test or assessment. We made efforts to mitigate these perceptions. For instance, previous studies have distributed activities in individual envelopes. In initial planning, we found this to be well outside the norm of classroom activities in the current setting, and instead used a collated packet of papers with a cover sheet to mask differences across conditions.

Some teachers also reported questions from students along the lines of "If this isn't graded, why do I have to do it?" One response was for teachers to justify the activities as part of a research study. Recognizing the potential for such deviations from instructions, researchers never described the project to teachers in terms of stereotypes, identity, or self-affirmation. Instead, researchers emphasized that the study concerned the thoughts and opinions of middle school students. Therefore, to the extent that teachers presented or justified the activities as part of a research project, they communicated that students' responses were valued, which we expected would encourage expressive self-reflection.

Comparison to Original Study

In the context of replication, it is important to be clear about key similarities and differences in protocol, subjects, and context. This is particularly true for interventions in applied school settings, where procedures must be sensitive to local conditions and can shift over time due to logistical constraints or contextual appropriateness. Previous self-affirmation interventions highlight this point: Sherman et al. (2013) reported creating simplified versions in a setting with many English Language learners, and even in the original setting, the experimental protocols (including the number of exercises, and instructions for choosing important values) shifted between years (Cohen et al., 2006).

The current study set out to replicate the original research (i.e., Cohen et al., 2006) as closely as possible at a larger scale in a new setting. Intervention materials—student exercises and teacher implementation instructions—were provided by the original research team. The fielded activities correspond most closely to Experiment

2 reported by Cohen et al. (2006)—circling important values instead of marking most and least important—and the simplified version employed by Sherman et al. (2013). Timing followed the original experiments, prioritizing a first administration as early in the school year as possible and spacing additional implementations throughout potentially stressful periods later in the school year.

The original study included three to five 7th grade implementations, depending on experiment (Cohen et al., 2009); we fielded three or four (depending on school) in both cohorts. In contrast to the original studies, we did not field implementations in 8th grade; a maximum of four implementations was feasible in the current context, and we prioritized the earliest activities. The original study also administered a student survey at the beginning and end of the 7th grade academic year. The survey addressed students' "self-perceived ability to fit in and succeed in school" (Cohen et al., 2009, p. 401). We conducted a similar survey at the beginning and end of the 7th grade school year for Cohort 2. In this respect, the Cohort 2 study was more similar to the original research than Cohort 1, when no surveys were administered.

The original study was conducted in a single school, described as "middle- to lower-middle-class families at a suburban northeastern middle school whose student body was divided almost evenly between African Americans and European Americans" (Cohen et al., 2006, p. 1307). The current context included students in 11 Midwestern middle schools in a single district. Overall student 7th grade enrollments in the district were 45% White, 25% Black, 19% Hispanic, and 10% Asian. Based on the original finding that results were consistent when non-Asian minority students were combined as "potentially stereotyped," we combined Black and Hispanic (including multiracial) students in preferred analyses. Across the 11 schools, the share of potentially threatened students ranged from 19% to 81%. As in the original study, the intervention was provided to students independently by teachers in their classrooms, with materials provided by the research team. The original study was conducted with 3 teachers. The current study (Cohort 2) was conducted with 44 teachers in 77 classrooms.

Our analyses include only administrative outcomes. It was not feasible to collect the more detailed outcome measures of the original study, including teacher gradebooks and a race activation task at the end of Grade 8 (Experiment 2) or Grade 7 (Experiment 1). However, we collected state standardized achievement test results, which were not considered in the original research.

Fidelity

Previous research provides little specific guidance on how to identify or measure the most relevant aspects of self-affirmation implementation, but the anecdotal challenges that teachers reported in implementing the activities in their classrooms highlight the need for more attention to these issues in applied settings. We considered several indicators of fidelity. One indicator is whether students responded to the writing prompts. By that standard, fidelity was quite high in both Cohort 1 and Cohort 2. In terms of basic exposure to the assigned materials, 88%–95% of students completed the assigned activity for each administration. Student absences from class accounted for the majority of noncompletion, whereas less than 1% of students in each administration completed a nonassigned packet due to administrative errors (such as a roster change).

We also coded the content of all students' responses, distinguishing between responses that showed clear evidence of self-affirming reflection and those that did not. Each response was coded independently by two trained coders who were blind to the experimental condition. A response was coded as self-affirming if it met three criteria: (a) the student wrote about themselves, (b) the response identified a listed "value" from the experimental prompt, and (c) the text expressed either the importance of the value (, e.g., "My family is the most important thing to me because . . .") or that they are "good in" the valued domain (example: "I'm good at drawing."). Interrater agreement was above 80% in both cohorts, and discrepant cases were resolved with the guidance of a core research team member. Based on those measures, fidelity to treatment was high in both cohorts, with 98.0% of treatment students providing at least one response reflecting self-affirming reflection, and 95.8% doing so during the first two exercises of the year.

Although our study is unprecedented in the scale at which we have documented fidelity in self-affirmation writing exercises, we acknowledge that it is possible for more subtle aspects of implementation to have failed in ways that we could or did not observe. Teachers' independent actions in the classroom, as discussed above, provide one example. Educational research has highlighted the organizational mechanisms that buffer teachers' practice from external demands (Weick, 1976) and the role of individual teachers' sense-making in shaping how reforms are enacted in the classroom (Coburn, 2004). We therefore gathered additional evidence with a teacher survey conducted at the end of each school year. These responses should be interpreted with caution for several reasons: we obtained reports from the teachers of only 56.0% of students (46.1% for Cohort 1 and 64.2% for Cohort 2), the items were retrospective reports (6 months on average after the fact), and it is unknown whether these (or any) teacher behaviors are critical to self-affirmation success. Nevertheless, these data complement other implementation measures and provide a preliminary window into teachers' administration of the activities.

Teacher responses supported the anecdotal reports discussed above, suggesting that the presentation of the exercises was not always as directed. Teachers of 31.1% of students reported describing the writing exercises as being part of a research study, and teachers of 20.3% of students reported describing the activities as "good for" students. These deviations may have detracted from the effectiveness of the self-affirmation activities, but we do not know how they compare to previous studies, because prior research has not reported systematically on teacher administration.

Sample

Because the study was administered in regular classrooms, all students in these classrooms completed some form of individual activity during implementation. However, students were only participants in the study (i.e., they were randomized to experimental condition, had data collected, and were included in analyses) if they assented and their parents consented. All seventh grade students in all 11 regular middle schools in the Midwestern school district were recruited to participate at school registration days (attended by the vast majority of parents and students) at the end of summer and with follow-up at the start of the school year. In the Cohort 1 study, we received consent and assent for 63.6% (1048/

1648) of the population; for Cohort 2 the number was 72.8% (1269/1722), reflecting improved recruiting efforts. Study participants were individually randomly assigned to the experimental group with randomization blocked by school.

Because attrition was low, even into 8th grade, we analyzed a consistent full cases sample. We dropped 9.0% of cases overall due to missing data/attrition: 2.6% of cases were missing data on covariates we included in models for precision, an additional 4.4% had no transcript data in 8th grade, and 2.1% more were missing standardized testing outcomes. The extent of attrition overall and the individual sources of attrition were statistically equivalent across experimental condition (Cohort 1: 10.6% treatment and 10.2% control, $\chi^2 = 0.03$, $df = 1$, $p = .86$; Cohort 2: 7.5% and 8.1% attrition, respectively, $\chi^2 = 0.14$, $df = 1$, $p = .71$); overall attrition was higher for Cohort 1 than Cohort 2 (10.4% vs. 7.8%, $\chi^2 = 4.75$, $df = 1$, $p = .03$). To the extent that differential attrition contributed to possible differences between cohorts, it would have operated (along with differences in recruiting) through different types of individuals being included in the two analytic samples, which we addressed explicitly (see "Individual Student Differences" Results section).

Measures

All student demographic information was derived from district administrative records. Our primary individual demographic variable was an indicator for students' potential susceptibility to social identity threats relating to academic performance in school, which we operationalized as African American or Hispanic racial/ethnic group membership. We treated multiracial students as potentially susceptible to racial identity threat because they are likely to identify with or be perceived as a member of a marginalized group, but results were similar when these students were excluded (see Figure 3, Panel C). To the extent that administrative racial/ethnic group membership misrepresents susceptibility to social identity threats, our impact estimates may have been attenuated, but similarly so for both cohorts.

To increase the precision of the self-affirmation treatment effect estimates, we included additional baseline student characteristics in our preferred specification for impact models. These included pretreatment (Grade 6) achievement outcomes and binary indicators for female, limited English proficiency status, receipt of special education services, and eligibility for free or reduced price lunch, which we included as a proxy for family economic resources. Results were substantively similar when we excluded these covariates (see Figure 3 and, in the online supplemental materials, Appendix Figure A1).

In some models, we restricted the sample to schools with relatively low proportions of Black and Hispanic students and relatively large prior achievement gaps for those students, both of which serve as proxies for more potentially threatening school contexts. Following previous research, we created a binary indicator for potentially threatening school contexts, defined as schools with below average numbers of Black and Hispanic students and above average prior racial achievement gaps (Hanselman et al., 2014).

Our ultimate interest was students' academic performance. The primary outcomes, following previous research in the self-affirmation literature, were students' overall GPA in Grade 7 and Grade 8. GPA reflects overall academic performance across all academic subjects and was recorded on a 4-point scale. Results were robust to focusing on only core academic courses, which corresponded closely to overall GPA (correlations of 0.98–0.99 in each grade). We gave Grade 8 GPA conceptual priority, as it was the only GPA measured entirely subsequent to the full treatment regime.

In supplementary analyses, we assessed treatment effects on a standardized academic assessment, the Wisconsin Knowledge and Concepts Examination (WKCE) tests in mathematics and reading. During the study period, WKCE tests were administered for state accountability purposes in November of Grade 7 and Grade 8. Although the Grade 7 tests were administered relatively early in the course of the intervention, the second exercise explicitly targeted the potentially high stress week prior to WKCE testing, making effects on this early outcome worthy of consideration.

Experimental Balance

Table 1 reports descriptive statistics and tests of baseline experimental equivalence for each cohort, both overall and within the subset of potentially threatened (Black and Hispanic) students. The sample was majority White, but included a substantial number of potentially threatened students in each cohort (reported numbers include multiracial students). Pretreatment differences between the treatment and control group were substantively small (generally less than 0.1 standard deviations) and not statistically significantly different, suggesting that randomization was successful in yielding comparable groups.

Analyses

All analyses were based on intention-to-treat estimates of the effect of self-affirmation, which assess the impact of assignment to the treatment group and therefore reflect the policy-relevant impacts of providing the self-affirmation (Borman, 2002). We calculated effects overall and within theoretically relevant subgroups. Estimates were based on the following general multilevel model of treatment effects:

$$Y_{ij} = \beta_0 + \beta_1(Treatment_i) + \beta X_i + \eta_j + \varepsilon_i \quad (1)$$

In this model, Y_{ij} is the observed outcome for student i in school j, $Treatment_i$ is the randomly assigned self-affirmation treatment status for student i, X_i is a vector of pretreatment covariates (Grade 6 outcome, gender, limited English proficiency, special education, and free lunch eligibility), η_j is the residual component for school j, and ε_i is the residual for student i. Because the treatment was randomly assigned to each student, β_1 provides an unbiased estimate of the effect of the self-affirmation intervention without additional controls, but we included a pretreatment achievement measure and additional covariates, X_i, to increase the precision of this estimate.[2]

[2] Some previous research has highlighted self-affirmation effects on achievement trajectories. These trends are especially helpful in characterizing the decline of minority students' achievement relative to majority students. We focus only on impacts on outcomes at single points in time here for two reasons: (a) our substantive interest is (variability in) the ultimate benefits of the intervention among potentially threatened students, which is best captured by overall impacts, and (b) given baseline equivalence, impacts on overall outcomes are analogous to impacts on (linear) trends. Estimates from longitudinal growth models were substantively similar to those presented here but less precise.

Table 1
Descriptive Statistics and Experimental Balance by Study, Overall and for Potentially Threatened Students (Black/Hispanic)

Variable	Cohort 1					Cohort 2				
	M	C *M*	T *M*	Std. diff. (C-T)	*p*	*M*	C *M*	T *M*	Std. diff. (C-T)	*p*
All students	[939]	[465]	[474]			[1,170]	[580]	[590]		
Female	.502	.520	.483	.075	.253	.499	.498	.500	−.003	.953
Potentially threatened	.353	.357	.348	.019	.776	.384	.367	.400	−.067	.250
American Indian	.039	.047	.032	.080	.218	.032	.028	.036	−.046	.434
Asian	.106	.092	.120	−.090	.168	.142	.147	.137	.027	.650
Black	.183	.163	.203	−.101	.122	.230	.209	.251	−.100	.086
White	.757	.768	.747	.049	.456	.702	.712	.692	.045	.443
Limited English proficiency	.144	.159	.129	.087	.184	.170	.167	.173	−.015	.798
Free/reduced lunch	.411	.413	.409	.007	.910	.463	.459	.468	−.018	.753
Grade 6 GPA	3.27 (0.64)	3.28 (0.65)	3.27 (0.63)	.009	.896	3.19 (0.67)	3.21 (0.68)	3.18 (0.67)	.042	.477
Grade 6 WKCE Math	525.3 (57.5)	522.2 (57.8)	528.4 (57.1)	−.108	.098	516.8 (51.7)	515.0 (51.1)	518.6 (52.3)	−.071	.227
Grade 6 WKCE Reading	510.8 (56.4)	508.0 (56.7)	513.6 (56.0)	−.100	.127	504.8 (57.1)	505.0 (57.4)	504.5 (56.9)	.009	.872
Black/Hispanic Students	[331]	[166]	[165]			[449]	[213]	[236]		
Female	.489	.512	.467	.091	.410	.566	.568	.564	.009	.923
Potentially threatened	1	1	1			1	1	1		
American Indian	.112	.133	.091	.132	.231	.082	.075	.089	−.050	.595
Asian	.009	.006	.012	−.064	.560	.020	.028	.013	.110	.244
Black	.520	.458	.582	−.248	.024	.599	.568	.627	−.120	.203
White	.568	.584	.552	.066	.548	.519	.521	.517	.008	.930
Limited English proficiency	.293	.343	.242	.221	.044	.294	.300	.288	.027	.775
Free/reduced lunch	.801	.819	.782	.094	.395	.851	.864	.839	.070	.461
Grade 6 GPA	2.85 (0.65)	2.83 (2.83)	2.87 (0.61)	−.061	.583	2.78 (0.65)	2.75 (0.63)	2.80 (0.66)	−.076	.420
Grade 6 WKCE Math	491.3 (53.1)	488.9 (55.2)	493.8 (51.1)	−.092	.406	486.1 (44.7)	482.6 (44.7)	489.3 (44.6)	−.149	.114
Grade 6 WKCE Reading	477.9 (53.3)	475.9 (51.9)	480.0 (54.8)	−.076	.490	471.9 (52.1)	471.3 (52.5)	472.4 (51.9)	−.021	.823

Note. Racial/ethnic indicators are not mutually exclusive and do not sum to 1 across groups. This table includes multiracial and White Hispanic students with potentially threatened students, as in our main specifications. Standard deviations in parentheses; sample sizes in brackets. WKCE = Wisconsin Knowledge and Concepts Examination; T = treatment; C = control; Std. Diff. = treatment-control in standardized units; p = p value for test of the null hypothesis that the difference (C-T) is equal to zero.

Within this basic framework, we conducted specific analyses to explore potential differences between the two studies, including alternate outcomes and estimates for theoretically relevant subgroups. Many of our analyses tested for differences in effects between Cohort 1 and Cohort 2 by estimating cohort-by-treatment interactions in pooled models with all observations, and we also estimated overall effects with the pooled data. We provide additional details for specific analyses as we present the results below.

Results

Estimated Impacts of Self-Affirmation

The raw pattern of results for the new study of self-affirmation (Cohort 2) for the focal outcome (GPA) is presented in the right panel of Figure 2. As expected, there were no effects of the intervention on the performance of Asian and White students, who are not hypothesized to be subject to the same types of identity threats in school as are the other groups. Potentially threatened groups (Black and Hispanic) performed worse overall, but the differences between treatment and control groups were similarly small in both 7th and 8th grade. To estimate treatment effects as precisely as possible for this targeted group, we used multilevel models of the self-affirmation intervention, controlling for pretreatment student characteristics. Estimates for all outcomes were negative, but none were statistically different from zero (see Table 2). The GPA effect in Grade 7 was approximately zero (d = −0.002), and the effect in Grade 8 was nominally negative (d = −0.072). Because the sample was quite large, these null results rule out (at the 0.05 significance level) impacts of 0.10 standard deviations or greater on GPA in Grades 7 and 8.[3] Results for standardized achievement outcomes were similar. Concerning our first research question, therefore, we found no evidence of treatment benefits for the targeted population in the new study.

Although not our primary focus, we also tested three additional findings reported by Cohen et al. (2006). First, we found no evidence of greater benefits of the intervention for potentially threatened students; the estimated interaction pointed in the opposite direction in our preferred specification but was not significantly different from zero (p = .15). Second, we found no evidence of differential effectiveness by prior academic performance. Following the procedures described by Cohen et al. (2006), we created tercile groups based on 6th grade GPA, within the potentially threatened and potentially nonthreatened groups. We failed to reject the null hypothesis that treatment impacts were equivalent across all three groups (p = .20). We also found no evidence of differential impacts by prior achievement among White and Asian students (p = .73). Finally, we tested for evidence of an improved

[3] The 95% confidence interval for the self-affirmation effect on overall GPA in Grade 7 was [−0.047, 0.165] for Cohort 1 and [−0.088, 0.083] for Cohort 2. The intervals for Grade 8 were [0.015, 0.282] and [−0.192, 0.047].

trajectory of performance throughout the year. Considering students' grades in each of the four terms of the school year, we tested for an interaction between treatment and term. GPA declined by 0.05 GPA points per term on average among Black and Hispanic students, but there was no difference by experimental condition ($p = .77$).

Comparing Self-Affirmation Effects Across Studies

The results above led us to ask whether the null effects in the current study (Cohort 2) differed from those in the previous research in the same setting (Cohort 1). A first question was whether the benefits observed previously (Borman et al., 2016) were detectable in the year following the intervention. We analyzed data from the subset of students from the prior study with valid observations in Grade 8, using parallel procedures to those above (estimates summarized in Table 2).[4] We found that self-affirmation group students received significantly higher grades in 8th grade ($d = 0.152$), bolstering the interpretation that the intervention led to detectable increases in academic performance for African American and Hispanic students. However, when we combined cases across studies, we did not find a statistically significant average self-affirmation treatment effect (Grade 7: $p = .54$, Grade 8: $p = .58$).

To address our second research question, we estimated the difference between self-affirmation impacts for Cohort 1 and Cohort 2 by pooling data from both samples and including cohort interactions with all covariates. We found that in several cases the null effects for Cohort 2 were distinguishable from comparable effects for Cohort 1. For the primary outcome, 8th Grade GPA, the standardized Cohort 2 estimate was small and negative ($d = -0.072$), whereas the Cohort 1 estimate was positive

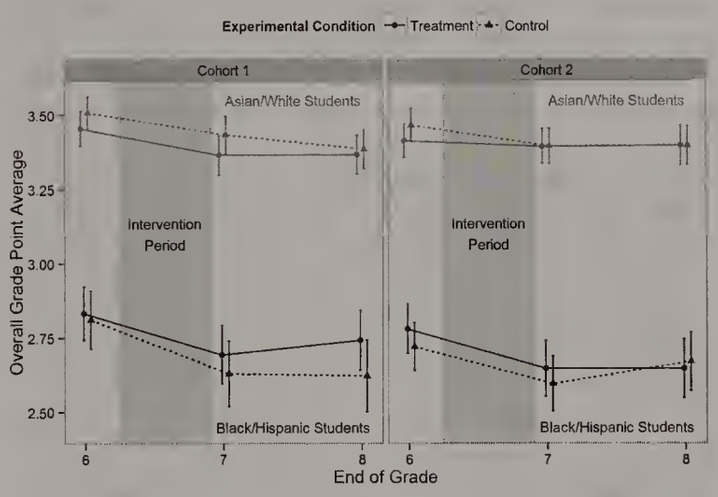


Figure 2. Yearly grade point average (with 95% confidence intervals) by race/ethnicity and experimental condition. Randomly assigned self-affirmation writing interventions were administered throughout the 7th grade year. No effects of the treatment are hypothesized for Asian and White students, who are not subject to general negative stereotypes about academic ability. Raw treatment versus control differences are statistically different from zero only for Black and Hispanic students in Grade 8 in Cohort 1. The treatment benefits in that cohort are statistically different than the small negative effect observed in Cohort 2. See Table 2 for standardized estimates and Table A3 in the online supplemental materials for results from a pooled treatment effects model.

Table 2
Standardized Self-Affirmation Treatment Impact Estimates for Black and Hispanic Students

Outcome	Cohort 1 ($N = 331$) Estimate	*SE*	Cohort 2 ($N = 449$) Estimate	*SE*	*p* value for difference
GPA, Grade 7	.062	.057	−.002	.043	.363
GPA, Grade 8	.152	.070	−.072	.058	.013
WKCE Mathematics, Grade 7	.072	.059	−.085	.047	.037
WKCE Mathematics, Grade 8	.101	.070	−.080	.044	.023
WKCE Reading, Grade 7	−.034	.069	−.005	.055	.737
WKCE Reading, Grade 8	−.030	.071	−.005	.056	.781

Note. All estimates are based on models including controls for pretreatment measures of the outcome and baseline student characteristics (gender, special education status, Limited English proficiency designation, and eligibility for free or reduced price lunch). See Table A3 in the online supplemental material for full pooled model results.*SE* = Standard Error; GPA = Overall grade point average; WKCE = Wisconsin Knowledge and Concepts Examination.

($d = 0.152$), and we could reject the null hypothesis that effects were equal ($p = .013$).[5] We also found statistical evidence of differences between the treatment effects across cohorts for the two supplementary mathematics state test score outcomes ($p = .037$ in Grade 7, $p = .023$ in Grade 8), although only the Grade 8 mathematics cohort effect difference would be statistically significant if the Bonferroni correction for multiple comparisons was applied to both estimates in this mathematics domain.

These results were robust across different specifications of the treatment effects model. In addition to our preferred specification, which included the full set of individual control variables, we also estimated impacts in models with no covariates and with controls only for the pretreatment outcome measure. Figure 3 summarizes results of these three specifications (represented by symbol shapes) for the focal group and comparison (Black/Hispanic students, combined control; Panel B1), as well as for alternate comparisons testing theorized moderators (discussed in the corresponding sections below). Appendix Figure A1 in the online supplemental materials presents comparable results for Grade 7 overall GPA. In all cases, results were substantively robust across all covariate specifications, although predictably less precise for the models omitting the alternate control cases.

To summarize results to this point, the two studies provided diverging pictures of the impacts of the self-affirmation interven-

[4] These analyses differed from previous reported by considering only students with Grade 8 information for all outcomes. The main implication was that the reanalyzed results were less precise, and therefore provided more conservative tests of statistical significance. The pattern of results across Grade 7 matched those reported by Borman et al. (2016)—positive benefits for GPA and mathematics achievement and smaller negative impacts on reading— although none of these were statistically significant in the reduced sample (see Table 2).

[5] Appendix Table A3 in the online supplemental materials presents all estimates from pooled models of treatment effects in both cohorts. These models suggest general similarity between cohorts in the associations between covariates and outcomes (fewer significant interactions than would be expected by chance). There is also suggestive evidence that the control group was higher achieving in Cohort 2 in GPA and mathematics, conditional on Grade 6 scores, but none of these differences are significant at the 0.05 level.

tion on Black and Hispanic students' academic outcomes. For Cohort 1, benefits in GPA persisted in the academic year following the intervention. For Cohort 2, however, we found no evidence of benefits of the intervention. Moreover, we rejected the null hypothesis that impacts were equal in both studies, despite being conducted in the same research setting. These results motivated our final research question: do the currently theorized moderators of self-affirmation explain the differences in treatment effects across the two cohorts? In the remaining sections, we focus on the primary outcome measure, Grade 8 GPA, and assess potential explanations for the decline in treatment effects from Cohort 1 to Cohort 2.

Differences in the Delivery of Self-Affirmation: Intervention Design

Research projects, like educational practice, evolve over time for pragmatic reasons. For instance, in previous self-affirmation studies, investigators adjusted the frequency and content of intervention exercises as they were implemented across successive cohorts and in new settings (Cohen et al., 2009; Sherman et al., 2013). In the current study, two design changes between the first and second cohort created differences in the delivery of the self-affirmation activities that potentially explain differential impacts: a shift in comparison group activities for one of the four exercises and a preintervention survey, which was added in the second study.

First, a randomly selected half of the control group was assigned a different first exercise in the Cohort 2 study, compared with Cohort 1. All control students were assigned the original control activity in Cohort 1, which directed students to select values that were unimportant to them and write about why these values may be important to someone else. Half of the control group did the same in Cohort 2, but half was randomly assigned to an alternate control activity for exercise 1 that asked students to write about what they did over the summer. Alternate control conditions were added in response to reported struggles of some students with the original "least important values" control activity. The alternate control writing prompt was modeled after typical classroom free-writing prompts, and was administered to nonconsented students in both years. This prompt is "neutral" in the sense that it does not explicitly refer to values, but students could, potentially, write self-affirming responses (see "Student Experiences" section below). A random half of the control group in both cohorts completed a comparable alternate activity for exercise 2, which asked students to describe how to complete a procedural task, such as how to open a locker.

To assess whether this modification in the control regime contributed to different intervention impacts, we focused on the randomly selected half of the control group in both cohorts that received exactly the same sequence of exercises, which directly followed the original design (Cohen et al., 2006). These estimates are presented in Figure 3 in subpanel 2 for each sample (labeled "Original Control"). The cohort-by-treatment interaction estimates were substantively unchanged in these analyses, though less precise owing to the smaller sample size, implying that the slight procedural change does not explain the drop-off in impact in the second cohort. Because we found no evidence of differences between the two control groups, we pooled both groups for all reported analyses, unless noted otherwise.

A second design change for the second cohort was the administration of a 15–20 min survey by researchers in classrooms in the first week of school. Interaction with research team members was similar for both studies because, for Cohort 1, researchers visited classrooms during this time to collect student assent forms. In both assent (Cohort 1) and survey (Cohort 2), researchers did not connect these overt research activities with the writing exercises, the first of which was administered on average 1 week later. Students were told in both cases that the study was interested in their thoughts and opinions as middle school students. The survey included items about individual characteristics (e.g., locus of control, self-complexity, and social belonging) but omitted any specific reference to racial identity, stereotypes, or self-affirmation, which might have primed students to experience identity threats.

It is theoretically possible that survey prompts about social–psychological constructs like social belonging could change how students respond to the self-affirmation exercises. Although we could not directly assess whether the inclusion of the survey accounted for lower benefits for Cohort 2, this explanation is unlikely for two reasons. First, to explain the decline in our setting, prior surveys would needed to have muted the treatment contrast (such as by inoculating treatment students from self-affirmation benefits), but the original large and persisting impacts were found in the presence of a presurvey (Cohen et al., 2009). Based on this result, we might have expected the largest benefits for Cohort 2. Second, the prior surveys were distinct from the self-affirmation exercises, fielded on a different day by the researchers, instead of teachers, and not explicitly linked to the exercises. Therefore social psychological responses activated by the survey would have to persist over time and remain relevant for a separate task. Although future research is necessary to test whether such prior prompts modify self-affirmation benefits, we note that if such brief, distinct stimuli moderate self-affirmation impacts, then there are many other school experiences that are also likely to matter. If true, the effects of the self-affirmation intervention would be extremely difficult to predict a priori.

Differences in the Delivery of Self-Affirmation: Student Experiences

One potential explanation for heterogeneity in treatment effects between the two studies is a decline in the quality of students' experience of the activities related to implementation. Although formal and informal procedures were consistent, the hypothesized psychological processes may be sensitive to subtle changes in delivery (Yeager & Walton, 2011), and it is possible that small changes in classroom procedures had large consequences for effectiveness. For instance, if teachers presented the materials differently in the second cohort, then fewer students may have engaged in genuine self-reflection. As discussed in the "Fidelity" section, no direct observations of classroom implementation were collected (the activities were intended to be part of regular classroom activities and not to be associated with research). Instead we conducted three indirect tests of implementation differences as explanations for differential benefits between cohorts: changes in theorized features of implementation, changes in implementing

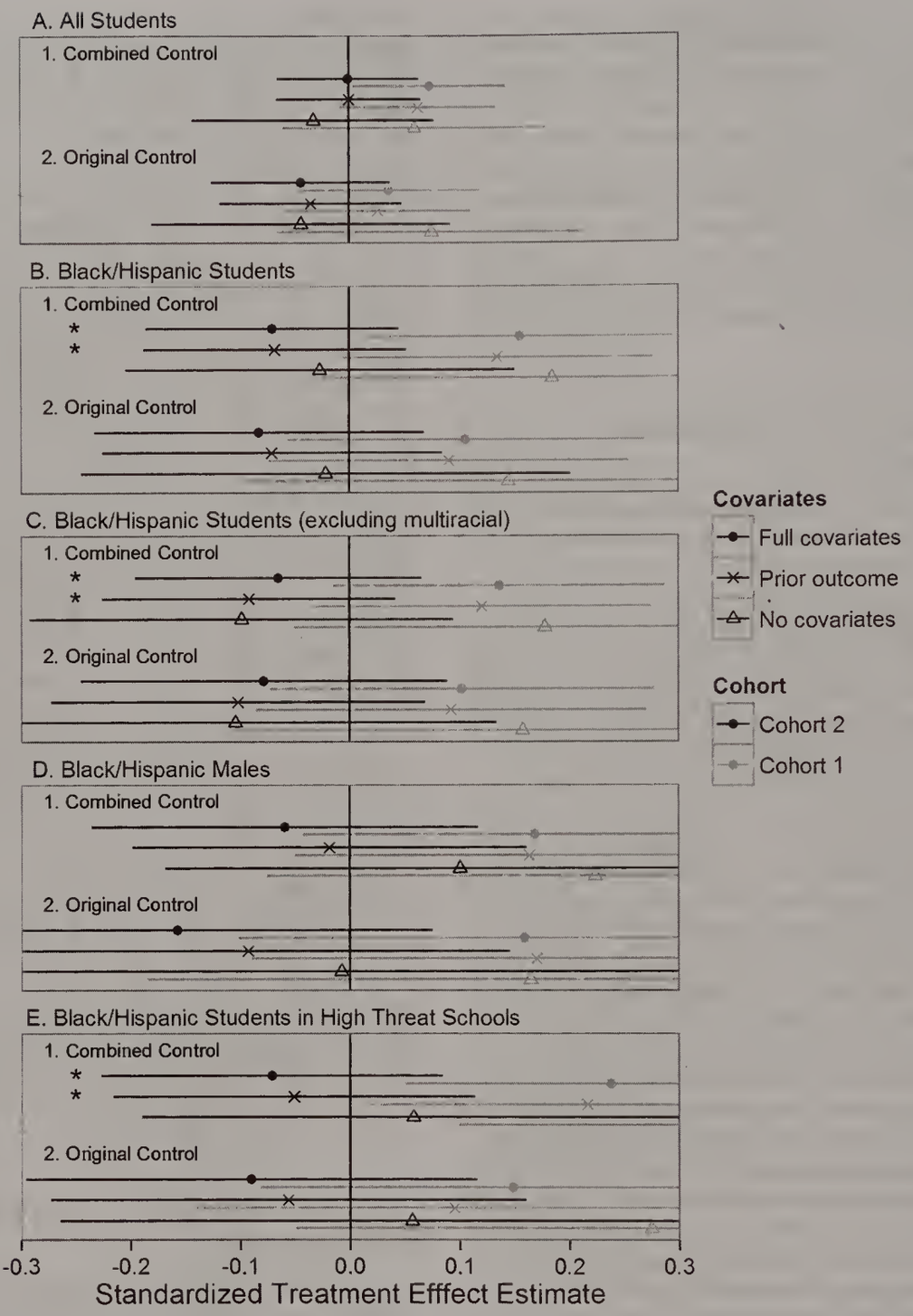


Figure 3. Estimated self-affirmation treatment effects on Grade 8 grade point average (GPA) by cohort, sample, comparison group, and included covariates. Each estimate was calculated from a separate multilevel model (students nested within schools) of intention to treat effect of the self-affirmation writing activities. Full covariates specifications include: Grade 6 GPA, gender, special education status, Limited English proficiency designation, and eligibility for free or reduced price lunch. Prior outcome is Grade 6 GPA. In the "Original Control" condition, students wrote about a least important value in each of the first two interventions. The "Combined Control" group includes these students as well as those who were assigned at least one writing prompt that did not explicitly mention values. For readability, the displayed range is restricted to effect sizes of absolute value 0.3 or less. Asterisks indicate that the estimated effects are statistically significantly different between cohorts ($p < .05$), based on a pooled model. The primary result, reported in Table 2, is the estimate for Black/Hispanic sample with combined control condition and full covariates (Panel B1 circles). Other results assess whether patterns were different for subpopulations and comparisons where self-affirmation benefits are hypothesized to be stronger and more consistent, as described in the text. Because the cohort difference persists across all specifications (although less precise in smaller subsamples), these tests provide no evidence that hypothesized moderators explain the difference.

teachers, and changes in students' written responses to the intervention.

First, we noted three theoretically important features of the self-affirmation writing intervention design: that activities are administered during targeted times of potential stress, especially early in the school year (Cook et al., 2012; Critcher et al., 2010), that activities are not explicitly presented as externally imposed (Silverman et al., 2013), and that activities are not presented as being beneficial to students (Sherman et al., 2009). We documented that that these features of implementation did not vary (or improved) between cohorts. With respect to timing, 91% of classrooms for Cohort 1 administered exercise 1 prior to the targeted first formative standardized assessment of the year, and 81% administered exercise 2 prior to the state standardized testing. The comparable numbers in Cohort 2 were 91% and 97%, respectively. Based on retrospective self-reports from teachers provided at the end of the school year, we also found more faithful implementation for the second cohort. In Cohort 1, 31.1% of students were taught by a teacher who reported describing the activities as "good for" them, whereas 42.2% were taught by a teacher who reported explaining the activities as connected to a research study. Both figures improved for Cohort 2: 13.9% for "good for" instructions and 24.6% for mention of a research study. With the caveats outlined in the "Fidelity" section, these reports show no indication of poorer implementation in Cohort 2. In other words, although imperfect delivery of the exercises may explain some of the attenuation of self-affirmation effects, these features did not explain the difference in effects between the two studies here.

Second, we considered whether changes in implementing teachers accounted for the decline in benefits. Due to staffing changes, 77% of the Black and Hispanic students in Cohort 1 and 60% in Cohort 2 completed the exercises with a teacher who implemented in both studies. If teacher fatigue with the study adversely affected implementation, then impact declines should have been largest among the "both-cohort" teachers. Conversely, if unique Cohort 1 teachers were especially effective, the declines should have been be largest among "single-cohort" teachers. We found no evidence for either hypothesis (see Appendix Table A4 in the online supplemental materials). Treatment by cohort interactions were substantively equivalent in both subpopulations (−0.196 grade points for the both-cohort teachers; −0.188 for the single-cohort teachers) and these interactions were statistically indistinguishable from one another ($p = .99$).

Finally, we tested whether students' written responses differed across the two cohorts of the study. Although features of the written responses are imperfect proxies for the desired self-reflection, they provide an indication of whether the quantity or quality differed across cohorts. The two most basic measures of overall engagement were comparable in both studies: exercise completion and words written. A high proportion of students completed the activities, ranging from 85–95% (Table A5, Column 1, in the online supplemental materials). Completion did not differ by experimental condition or cohort. In supplementary analyses, we found that completers tended to have higher prior GPA than noncompleters—no other baseline covariate predicted completion—but this difference was not distinguishable between cohorts.

The relative length of students' responses was consistent across cohorts too, after accounting for variation due to differences in prompts over time (Columns 2 and 3). The only treatment-control difference between cohorts was in mean words written for exercise 1 (Panel A), and this was completely explained by the randomly assigned "neutral" comparison group; students were more prolific when writing about their summer (in Cohort 2) than about an unimportant value. Comparing students with the same, "original" prompts (Column 3), there were no cohort differences. By these measures, basic engagement with the activities was consistent across the two cohorts.

Analyses of the qualitative measure of students' responses to the exercises (introduced in the "Fidelity" section above) implied that treatment caused students to engage in much higher rates of affirmation across all exercises in both studies.[6] The estimates are based on linear probability models, so the coefficient of 0.709 (Table A5, Panel B, Column 4, in the online supplemental materials) implies that the chance of affirmation writing was 71 percentage points higher in the treatment group in Cohort 1 for exercise 2. The interaction coefficient (0.0796) implies that this treatment effect was actually higher in the second cohort, at a significance level of $p < .1$. Exercise 1 was again an exception, but the difference was solely explained by the modifications to the control group (see Column 5). Not surprisingly, the control group in Cohort 2, including students who wrote about their summer, was more likely to write affirming statements, which others have noted is a risk in choosing that type of comparison activity (Cohen, Aronson, & Steele, 2000). Even so, treatment impacts on self-affirming writing were greater than 40 percentage points (0.427 = 0.721–0.294) in the second cohort overall.

On balance, analyses of implementation features, consistent teachers, and direct measures of intervention responses did not support the hypothesis that declines in implementation quality could explain lower benefits for Cohort 2. In particular, responses to the exercises were strong overall, and comparable between cohorts. These results cannot rule out the possibility of differential psychological responses to the exercises in the two implementations, which deserves attention in future research. However, for this possibility to be true, the association between key psychological responses and the desired features of students' written responses must have changed between cohorts. The more parsimonious explanation is that declines in implementation did not account for lower effectiveness.

Individual Student Differences

The success of social–psychological interventions depends fundamentally on individual characteristics. Self-affirmation is only hypothesized to help students who are subject to identity threat, and students may also differ in how they respond to the specific reflective writing activity. Meaningful individual differences between cohorts could have resulted from sampling variability and/or because the second cohort study sample was larger, including 36% more potentially threatened students (449 vs. 331 in Cohort 1), and different in terms of mean individual characteristics (see Table 1), due to more successful recruitment. We used three strategies to test for individual-level explanations of cohort differences: effects in theoretically sensitive subgroups, observable differences between

[6] Treatment effects are muted in exercise 3 for both cohorts because overall impacts include several schools that opted out of this exercise, and therefore students had no opportunity to engage in affirmation.

the two cohorts, and the plausible influence of unobserved heterogeneity.

One implication of theorized moderation of self-affirmation benefits by individual characteristics is that results should be consistently stronger, and therefore less variable across cohorts, in subpopulations where academic stereotype threats are hypothesized to be most salient. We tested effects in two such subpopulations: students identified as only Black or Hispanic (excluding multiracial students), who may identify more strongly with a stereotyped identity, and Black/Hispanic Males, who may be subject to the most acute general academic stereotypes in middle school (Purdie-Vaughns & Eibach, 2008). Results are summarized in Panels C and D of Figure 3. Contrary to the individual difference hypotheses, differential effects across cohorts were similar in both of these subpopulations, even though lower precision in the male subgroup led similar size differences to be statistically insignificant.

We also tested all observed individual student characteristics as explanations of cohort differences. For individual characteristics to explain the decline in treatment effects, differences between the two samples must have been related to treatment effect heterogeneity. We did find some descriptive differences between studies (see Table 1): the sample for Cohort 2 had more female students (52.6% vs. 49.8%; $p = .03$), lower 6th Grade GPAs on average (2.78 vs. 2.85; $p = .11$), and more students eligible for free or reduced price lunch (85.1% vs. 80.1%; $p = .07$). However, we found no statistically significant interaction between treatment and individual characteristics (Grade 6 GPA, gender, English proficiency, or Special Education designation) in either cohort, suggesting little opportunity for individual observed characteristics to explain different treatment effects. Not surprisingly, when we reweighted individual cases in each cohort to balance populations in terms of each of these observable characteristics (for instance, giving greater weight to poor students in Cohort 1, who were relatively underrepresented in that sample), the effect estimates in each cohort were substantively unchanged (see Table A7 in the online supplemental materials).

More generally, we gauged how large total (including unobservable) subpopulation differences would need to be to explain the different estimates between the two cohorts, assuming that individual-level treatment effects were constant over time. We considered a thought experiment in which the population was composed of two types of students: strong self-affirmation responders that benefit most from the intervention (Type A), and weak self-affirmation responders that benefit least (Type B). Assuming the boundary case that the Cohort 1 Black/Hispanic sample was populated solely by strong responders, then an estimate of the average impact for this type of student (d_A) on Grade 8 GPA is 0.152. Assume the Cohort 2 sample was comprised of a mixture of students of Type A and B, with the effects for Type B students (d_B) unknown. The total impact in Cohort 2 would then be an average of the two type-specific effects, weighted by the share of teach type (p_A and p_B, respectively):

$$d_{cohort\,2} = p_A(d_A) + p_B(d_B).$$

Based on the total effect estimate in Cohort 2 (−0.072) and the fact that the proportions of Type A and Type B students sum to 1, this implies:

$$d_{cohort\,2} = -0.072 = (1 - p_B)(0.152) + p_B(d_B)$$

Rearranging algebraically:

$$d_B = \frac{.224}{p_B} + (.152)$$

The implication of this inverse relationship between the share and effect size for weak-responders is that Cohort 2 null effects could only be explained by very large shares of weak-responders or by substantially negative effects for these students. For instance, if only the surplus students in Cohort 2 (25%) were weak responders, then the effect of the intervention among this population of students must have been −0.74 (= −.224/.25 + .152) to explain the total Cohort 2 impact; if half of the Cohort 2 population was the second type of student, then effects for this group would need to be −0.30 (= −.224/.5 + .152).[7] Because such drastic changes in the underlying population and such large negative effects of the intervention are not plausible, it is unlikely that differences in the underlying student populations explain cohort differences.

Changes in Social Context

Social–psychological interventions are also theoretically sensitive to features of the social environment in which they are implemented (Yeager & Walton, 2011). Because the studies for both cohorts were conducted in the same classrooms, schools, and district, we expected there to be relatively small differences in the relevant social conditions that students experienced across cohorts. This intuition was not directly testable, as there are no definitive measures of the relevant contextual features, but we assessed several indirect indicators of contexts that may be meaningful. We considered the demographic characteristics of the school population, differences in aggregate achievement, and school-specific impact estimates.

Previous research using data from the Cohort 1 study suggested that school contexts moderated the self-affirmation treatment effect on 7th grade outcomes, with the greatest benefits in schools with low minority populations and large prior achievement gaps (Hanselman et al., 2014). In new analyses (summarized in the Figure 3, Panel E), we found that larger than average treatment benefits in these schools in Cohort 1 persisted into 8th grade; however, self-affirmation benefits were no more consistent across cohorts in the population of "High Threat" schools, suggesting that context moderation does not explain the overall decline.

In addition, we considered whether shifts in demographic context of all students in the school (conceptually and empirically distinct from individual characteristics of the study samples discussed above) plausibly explained the difference in effects between cohorts. We found no evidence of this possibility, primarily because student characteristics did not change substantially between studies. One proxy for broad context differences related to academics and racial/ethnic identity is subgroup academic achievement and achievement gaps, which were similar for both cohorts and consistent with historic patterns (Figure A2 in the

[7] Similar calculations using the upper bound of the 95% confidence interval for the treatment effect in Cohort 2 results in necessary effects for the new student population of −0.29 as a 25% share of Cohort 2 and −0.07 as a 50% share.

online supplemental materials). At the school level, racial/ethnic cohort composition was similar in both cohorts, whereas achievement gaps, which are one proxy for a racialized academic school environment, were consistently large (Figure A3 in the online supplemental materials). Moreover, controlling for either school-level racial/ethnic composition or prior achievement gaps did not alter the core treatment-by-cohort interaction estimate, suggesting that these documented school characteristics did not account for the decline in treatment effects in the second cohort.

Finally, we estimated school-specific impacts for Black and Hispanic students using data from both cohorts to assess whether patterns were consistent across these local contexts. Effects in most schools were similar or slightly lower for the second cohort (Figure A4 in the online supplemental materials), suggesting general consistency in lower impacts in Cohort 2. However, dramatic changes from positive estimates for Cohort 1 to negative estimates for Cohort 2 were apparent in two schools (labeled points 5 and 11 in Figure A4 in the online supplemental materials). These differences may have been due to either drastic consequential changes in the local context or sampling variation. The latter is a more parsimonious explanation in light of the consistent demographic context discussed above, post hoc qualitative checks (which revealed no substantial year-to-year differences at these schools), and the implausibly large magnitude of the point estimate of the interaction for these schools (0.4–0.5 standard deviations).

To assess whether individual schools drove the overall results, we reestimated pooled treatment effect models omitting each of the 55 unique pairs of schools in the study (see Figure A5 in the online supplemental materials). The main results—small positive effects for Cohort 1, slightly negative effects for Cohort 2, and therefore a consequential interaction—held in all omitted samples. One school (11) stood out as an extreme case: Omitting this school reduced the interaction effect by 20%–40% (depending on which additional school was also omitted), whereas the range for all other omitted pairs estimates was within 15% of the overall estimate. Subsamples that excluded school 11 exhibited greater similarity in estimates across cohorts (smaller interactions) due mostly to smaller estimated benefits for Cohort 1, but also due to somewhat smaller estimated negative effects for Cohort 2. On the whole, although a single school contributed the most to the decline in effectiveness between cohorts, the differences were meaningfully large without it.

Classroom and district context features may also have contributed to the difference in treatment effects across cohorts. However, we did not have strong a priori predictions about the importance of features at either level. To the extent that individual teachers shape the relevant features of the classroom environment, the similarity in effects for consistent and inconsistent teacher populations (reported above) suggests a small role for these factors. At the district level, even substantial system-wide events are especially difficult to connect theoretically to differences in the treatment effect. For instance, there was notable political and civic unrest during the study surrounding legislation limiting public sector unions, rhetoric surrounding teachers' work, and school closures due to teacher protests. Schools in the district were closed for four days in February during the Cohort 1 study, and the associated gubernatorial recall election occurred in June between the two self-affirmation studies. We do not have strong theoretical predictions about whether these events translated to differences in school environments that moderated self-affirmation effects, but it seems unlikely that the unrest and missed days of regular schooling were critical to intervention success in Cohort 1. More generally, this example highlights that if self-affirmation effects are sensitive to context changes such as public debate about education then they are fundamentally fragile in the sense that relevant critical conditions are difficult to diagnose, and more importantly, to anticipate.

Discussion

The replication results reported in this article provide new evidence concerning two fundamental questions about the potential of self-affirmation interventions to improve academic performance and close achievement gaps (Cohen et al., 2006; Yeager & Walton, 2011): 1) Are there benefits of self-affirmation interventions for academic performance in middle school? and 2) Can we identify the necessary and sufficient preconditions for self-affirmation success? The large-scale replication results reported here, coupled with extensive post hoc tests of heterogeneous effects, provide disconfirming evidence on both counts: we found no effects of the intervention for Cohort 2, and we found no evidence that moderators from existing theory explained why this result differed from those in a previous study in the same setting. These results rule out important hypotheses about self-affirmation effects, both in terms of the magnitude of benefits and the sufficiency of theorized moderators, which refines our understanding of both fundamental questions. In closing, we elaborate these specific contributions, highlighting the unique evidence provided by this multicohort large-scale replication and implications for future research.

Are There Benefits of Self-Affirmation Interventions at Scale for Academic Performance in Middle School?

An important contribution of this article is that it reports on a new large-scale replication of the promising self-affirmation writing interventions introduced by Cohen et al. (2006). Comprehensive null results from this experiment provide no evidence of self-affirmation benefits, and the precision of the impact estimates rules out benefits that are as large as one third the size of those reported by Cohen et al. (2009). Like the recent replication by Dee (2015), our results suggest that self-affirmation has at best modest benefits for minority students when implemented at a large scale. Unlike that study, however, the current failure to replicate cannot be plausibly attributed to idiosyncratic features of the research site or procedures, because a similar prior replication in the same setting did find benefits (Borman et al., 2016). In this article, we reported persistent intervention benefits for the prior cohort and documented similarity in implementation measures across cohorts, including features of students' written responses.

It is important to point out that low statistical power is only a likely explanation for the null results in Cohort 2 if the true effect of the intervention was smaller than estimated for Cohort 1 and much smaller than in initial studies (Cohen et al., 2009; Sherman et al., 2013). Using the post hoc power calculations suggested by Gelman and Carlin (2014), we investigated the power of our Cohort 2 study design for a range of true effect sizes (see Figure 4). If the true benefit of self-affirmation on Grade 8 GPA was 0.30, similar to the initial study, then our power was above 0.99. If the

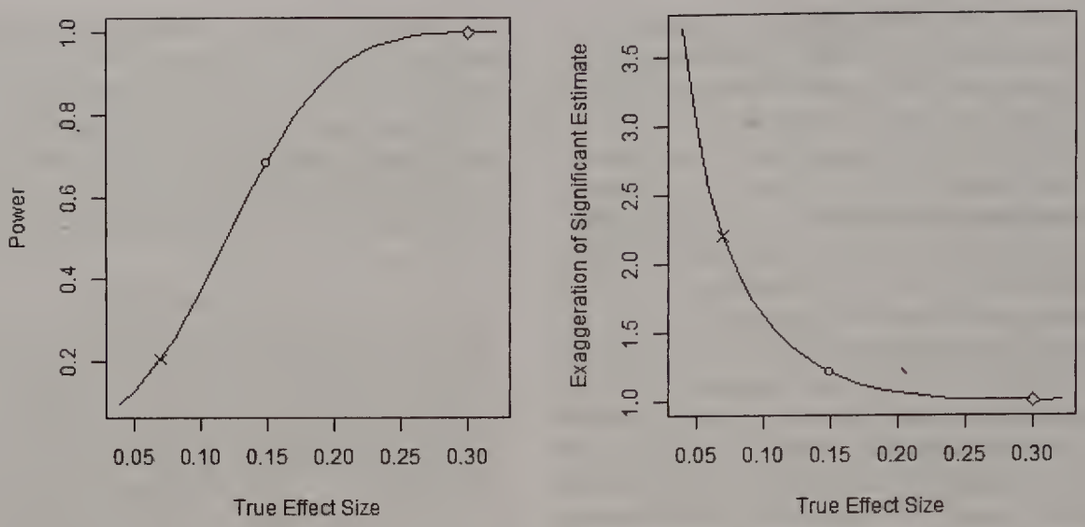


Figure 4. Power calculations for range of true effect sizes of self-affirmation intervention effects. Curves represent power (left panel) and expected exaggeration of a treatment effect estimate significant at the 0.05 level (right panel) for self-affirmation effects in Grade 8, given the design for new study (Cohort 2) reported here. Calculations are based on the procedure suggested by Gelman and Carlin (2014). Diamonds represent an effect size of 0.3, consistent with the initial study of self-affirmation interventions (Cohen et al., 2006); if true effects are this large, then power is virtually 1.0 and expected exaggeration is minimal. Circles represent the estimated effect size for the first cohort of students ($d = 0.15$). If the true effect were this large, then Cohort 2 power would be 0.68 and expected exaggeration would be 1.21. Xs represent the mean effect size calculated in Figure 1 ($d = 0.07$). If the true effect were this large, power would be 0.21 and significant values would exaggerate the true effect by 2.22 times on average.

true effect was 0.15, as estimated for Cohort 1, then power was 0.68. However, if the true effect size was 0.07, the average across the studies summarized in Figure 1, then this study had only a 21% chance of detecting an effect and a Type II inferential error was to be expected.

These power calculations highlight a more general possibility: the true impacts of these brief self-affirmation interventions may be positive but relatively small when implemented at scale and across heterogeneous contexts. As Bryk, Gomez, and Grunow (2011) observe, "the history of educational innovation is replete with stories that show how innovations work in the hands of a few, but lose effectiveness in the hands of the many" (p. 130; see also: Schneider & McDonald, 2006). This could be true for self-affirmation due to implementation challenges or differential effects across contexts. If so, then even very large field trials, such as the one conducted by Dee (2015) and the current study, are underpowered and unlikely to detect effects reliably. An important corollary implication, if the true effect size is small, is that significant estimates in individual trials are expected to overstate the magnitude of the effect by a substantial amount (Gelman & Carlin, 2014). If the true effect size is 0.07, then statistically significant results from the current design would overstate this effect by a factor of 2.2 in expectation.[8]

The plausible magnitude of self-affirmation effects is a crucial consideration for future work in this field, including implications for study design. If the true self-affirmation effect size for Black and Hispanic students when implemented on a large scale is 0.07, then we are aware of no studies with adequate power to reliably detect the effect, and statistically significant published results are likely to overstate the true impacts. The practical importance of such a small effect may be debatable, but from a policy perspective even a small benefit at scale could justify the negligible cost of this intervention. For instance, the benefits of the Tennessee STAR class size reduction experiment have been estimated to be 0.07 standard deviations in student reading achievement per $1,000 in per-pupil expenditure (Borman & Hewes, 2002, p. 258). A comparable benefit for brief self-affirmation activities, which are orders of magnitude less costly, would be very valuable for educators and policymakers. Therefore, more precise evidence about even potentially small effects of self-affirmation are needed. However, we recognize that more effective implementation of self-affirmation activities may be more expensive, especially if it requires dynamic guidance from a dedicated "psychological engineer" (Yeager & Walton, 2011). If this approach proved successful, then policy implications would then depend on the trade-off between greater benefits and costs.

Can We Identify the Necessary and Sufficient Preconditions for Self-Affirmation Success?

A second key contribution of this article is our detailed analysis of the differential effects of self-affirmation in two large-scale studies conducted in the same research setting. The results are puzzling in their lack of definitive explanation for differences, but they are informative because they demonstrate variation that cannot be explained by the moderators of self-affirmation benefits that

[8] Note that if the same scenario (true effect of 0.07) were true for the previous study (Cohort 1), then our results (estimated significant effect of 0.15) would make the correct inference about the existence of a positive effect but overstate the magnitude of this effect by approximately the amount expected by a significant effect for this study design.

have been proposed in the literature (see summary in Table 3). Our general conclusion is that the current hypotheses about variation in self-affirmation effects are insufficient to explain the potentially subtle moderators of impacts. We highlight three specific and related implications of the results.

First, our analyses demonstrate the value of tests of moderators to assess theory about where, and ultimately how, specific interventions are successful. The tests conducted here provide strong, if indirect, evidence about hypothesized differences due to implementation, individual, and context characteristics. Our assessment of individual differences is notable in this regard. Even though we did not directly measure all potential individual difference moderators, we calculated that the offsetting negative impacts of self-affirmation required for an individual difference moderator to explain the cohort differences were too large to be plausible. As a result, theorized differences in individuals across the two cohorts are unlikely to explain the heterogeneous results. In addition, our tests of moderators draw on the analytic leverage provided by a within-research site comparison across multiple cohorts and on the collection of relatively detailed implementation data, including students' written responses. This demonstrates the value of replication over time within a consistent research setting.

At the same time, unexplained variability highlights the need for additional inquiry into the implementation of these activities in diverse educational settings. Our attention to teachers' delivery of the activities and students' responses in large-scale implementations provides a first step in measuring variation in the implementation of self-affirmation exercises, but more work is needed to identify the necessary components for success. One insight from the scale-up effort reported here is the potential tension between fidelity to the scripted intervention and adaptation to local classrooms. At scale, teachers are unlikely to have close, long-standing relationships with researchers, and they are likely to respond to this tension in different ways. Some responses may have undercut the potency of the intervention, even though they did not preclude benefits in Cohort 1 and they did not seem to explain the different results in Cohort 2. One future direction could be to remove teachers from delivery through computerized implementation. However, the protocol might alternatively be modified to include teachers more fully. Our anecdotal interactions suggest that teachers would implement much more organically if they were allowed to read students' responses. Future research could explore implications for implementation and effectiveness.

Second, our results point to the need to develop the theory and evidence about how and where self-affirmation works. Because we tested a comprehensive list of proposed moderators of self-affirmation and failed to explain the variation in our findings between cohorts, we conclude that the current cadre of moderators offered by the literature is insufficient. Future experimental studies are needed to robustly assess the existing theorized moderators, and it may be that current theory needs to expand to incorporate new potential explanations for self-affirmation effects.

Our results call more attention to the overall lack of empirical evidence about moderators of self-affirmation effects, which makes it difficult to judge whether theory testing or expansion is the more crucial next step for the field. For example, there is little relevant data and few studies assessing whether awareness about the benefits of self-affirmation, one of the best substantiated po-

Table 3
Summary of Tested Hypotheses

Hypothesized explanation for difference in effects	Empirical tests of consistency between cohorts	Result
	Different effects due to features of the intervention delivery/implementation	
Providers	Consistent benefits for teachers implementing in both cohorts?	No
	All changes in benefits are due to teachers implementing in both cohorts (due to fatigue)?	No
Control group	Consistent benefits when compared to students in the original control condition?	No
Stealth	Teachers report more violations of protocol in second cohort: describing the activity as externally imposed research?	No
Awareness of Purported benefits	Teachers report more violations of protocol in second cohort: describing the activity as "good for you"?	No
Timing	Intervention more likely to miss key stressful periods in second cohort?	No
Engagement with the prompt	Students complete fewer exercises in second cohort?	No
	Students write fewer words in second cohort?	No
	Impact on self-affirming writing is different in second cohort?	No
	Different effects due to individual characteristics	
Racial group	Consistent benefits for all Black and Hispanic students?	No
	Consistent benefits for nonmultiracial Black and Hispanic students?	No
Race and gender	Consistent benefits for male minority students?	No
Prior achievement and other administrative characteristics	Consistent benefits when populations are re-weighted across cohorts on observable characteristics?	No
Unobserved receptivity to self-affirmation	Magnitude of different benefits for unobserved populations are plausible?	No
	Social context differences	
Broad (district) racial and academic climate	Different representation of racial minorities for the second cohort?	No
	Lower racial achievement differences for the second cohort?	No
School racial and academic climate	More consistent benefits in "high threat" schools with few minority students and large gaps?	No
	Differential benefits explained by one or two schools?	No

tential influences, moderates the effectiveness of the intervention. Sherman et al. (2009) is frequently cited for this point, but this article only shows a correlational relationship between awareness and affirmation effects on task performance. More research is needed to isolate to what extent this and other theorized components contribute to effectiveness.

Moreover, the unique challenges that arise at scale highlight the need for future research to consider the necessary and sufficient conditions of self-affirmation in applied settings. Our results point to two important avenues in future research: measures of features of implementation and variations in protocol. First, future research needs to develop systematic measures of implementation. This may include videos or observations of classrooms or, alternatively, getting more detailed information from classroom teachers soon after implementation in the form of interviews or surveys. Similarly, administrative data offer imperfect proxies for the social context in which self-affirmation takes place. School climate instruments, including measures of overt and subtle forms of bias and discrimination, should be tested as more direct indices of context. A stronger measurement component would allow researchers to assess how potentially relevant environmental changes, such as the political unrest that occurred during the research reported here, did or did not translate into differences in schools.

Another suggestion for future self-affirmation research in applied settings is to experiment with features of the delivery of the intervention. For instance, researchers might contrast computerized delivery (Paunesku et al., 2015), which may help standardize the delivery of the intervention, to delivery by classroom teachers who, alternatively, may play important roles if their students believe that the values-affirming exercises are coming from them. If teacher-based delivery is employed, our experiences suggest that teacher protocols are an important area to focus on, because even with a script individual teachers may implement materials differently. By systematically varying these protocols, future research should consider how different instructions affect the activities being presented as beneficial, and whether this explains differential benefits.

Third, our unexplained heterogeneity results imply practical limitations of self-affirmation as a tool to improve student performance and close achievement gaps. The proposed efficacy of brief social–psychological interventions to improve educational performance is specific, requiring tailoring the right kind of program to the right kind of students in the right kind of social environment (Walton, 2014; Yeager & Walton, 2011). If variability in impacts cannot be predicted with the information available to educators, then the practical value of these interventions is unclear. That said, short self-affirmation writing exercises in the classroom remain a virtually costless approach to potentially addressing some of the racial disparities in school. Students often participate in broadly similar writing activities in the classroom during the school day, and targeted self-affirmation activities are unlikely to negatively impact students. The impacts may well be positive, but they are likely small, and our results suggest that challenges remain in predicting where exactly, and therefore how widely, the potential benefits of self-affirmation writing activities will extend.

References

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35,* 29–46. http://dx.doi.org/10.1006/jesp.1998.1371

Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education, 77,* 7–27. http://dx.doi.org/10.1207/S15327930PJE7704_2

Borman, G. D., Grigg, J., & Hanselman, P. (2016). An effort to close achievement gaps at scale through self-affirmation. *Educational Evaluation and Policy Analysis, 38,* 21–42. http://dx.doi.org/10.3102/0162373715581709

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24,* 243–266. http://dx.doi.org/10.3102/01623737024004243

Bowen, N. K., Wegmann, K. M., & Webber, K. C. (2013). Enhancing a brief writing intervention to combat stereotype threat among middle-school students. *Journal of Educational Psychology, 105,* 427–435. http://dx.doi.org/10.1037/a0031177

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217–224. http://dx.doi.org/10.1016/j.jesp.2013.10.005

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M. T. Hallinan (Eds.), *Frontiers in sociology of education* (pp. 127–162). New York, NY: Springer.

Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education, 77,* 211–244. http://dx.doi.org/10.1177/003804070407700302

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin, 26,* 1151–1164. http://dx.doi.org/10.1177/01461672002611011

Cohen, G. L., & Garcia, J. (2008). Identity, belonging, and achievement: A model, interventions, implications. *Current Directions in Psychological Science, 17,* 365–369. http://dx.doi.org/10.1111/j.1467-8721.2008.00607.x

Cohen, G. L., & Garcia, J. (2014). Educational theory, practice, and policy and the wisdom of social psychology. *Policy Insights From the Behavioral and Brain Sciences, 1,* 13–20. http://dx.doi.org/10.1177/2372732214551559

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006, September 1). Reducing the racial achievement gap: A social–psychological intervention. *Science, 313,* 1307–1310. http://dx.doi.org/10.1126/science.1128317

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009, April 17). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324,* 400–403. http://dx.doi.org/10.1126/science.1170769

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology, 65,* 333–371. http://dx.doi.org/10.1146/annurev-psych-010213-115137

Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology, 102,* 479–496. http://dx.doi.org/10.1037/a0026312

Critcher, C. R., & Dunning, D. (2015). Self-affirmations provide a broader perspective on self-threat. *Personality and Social Psychology Bulletin, 41,* 3–18. http://dx.doi.org/10.1177/0146167214554956

Critcher, C. R., Dunning, D., & Armor, D. A. (2010). When self-affirmations reduce defensiveness: Timing is key. *Personality and Social Psychology Bulletin, 36,* 947–959. http://dx.doi.org/10.1177/0146167210369557

Dasgupta, N., Scircle, M. M., & Hunsinger, M. (2015). Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 4988–4993. http://dx.doi.org/10.1073/pnas.1422822112

Dee, T. S. (2015). Social identity and achievement gaps: Evidence from an affirmation intervention. *Journal of Research on Educational Effectiveness, 8,* 149–168. http://dx.doi.org/10.1080/19345747.2014.906009

Eagly, A. H., & Kite, M. E. (1987). Are stereotypes of nationalities applied to both women and men? *Journal of Personality and Social Psychology, 53,* 451–462. http://dx.doi.org/10.1037/0022-3514.53.3.451

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science, 9,* 641–651. http://dx.doi.org/10.1177/1745691614551642

Hanselman, P., Bruch, S. K., Gamoran, A., & Borman, G. D. (2014). Threat in context: School moderation of the impact of social identity threat on racial/ethnic achievement gaps. *Sociology of Education, 87,* 106–124. http://dx.doi.org/10.1177/0038040714525970

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen. C. J., Blair, S. S., Rouse, D. 1., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology, 106,* 375–389. http://dx.doi.org/10.1037/a0034679

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2015). Closing achievement gaps with a utility–value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspp0000075

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11,* 365–371. http://dx.doi.org/10.1111/1467-9280.00272

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645–654. http://dx.doi.org/10.1177/1745691612464056

Kost-Smith, L. E., Pollock, S. J., Finkelstein, N. D., Cohen, G. L., Ito, T. A., Miyake, A., . . . Singh, C. (2012). Replicating a self-affirmation intervention to address gender differences: Successes and challenges. *AIP Conference Proceedings, 1413,* 231–234. http://dx.doi.org/10.1063/1.3680037

Lauer, S., Momsen, J., Offerdahl, E., Kryjevskaia, M., Christensen, W., & Montplaisir, L. (2013). Stereotyped: Investigating gender in introductory science courses. *CBE Life Sciences Education, 12,* 30–38. http://dx.doi.org/10.1187/cbe.12-08-0133

McQueen, A., & Klein, W. M. P. (2006). Experimental manipulations of self-affirmation: A systematic review. *Self and Identity, 5,* 289–354. http://dx.doi.org/10.1080/15298860600805325

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science, 330,* 1234–1237. http://dx.doi.org/10.1126/science.1195996

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science, 18,* 879–885. http://dx.doi.org/10.1111/j.1467-9280.2007.01995.x

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536. http://dx.doi.org/10.1177/1745691612463401

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science, 26,* 784–793. http://dx.doi.org/10.1177/0956797615571017

Purdie-Vaughns, V., Cohen, G., Garcia, J., Sumner, R., Cook, J., & Apfel, N. (2009). Improving minority academic performance: How a values-affirmation intervention works. *Teachers College Record.* Retrieved from http://www.columbia.edu/cu/psychology/vpvaughns/assets/pdfs/Improving%20Minority%20Academic%20Performance%20(2009).pdf

Purdie-Vaughns, V., & Eibach, R. (2008). Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex Roles, 59,* 377–391. http://dx.doi.org/10.1007/s11199-008-9424-4

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115,* 336–356. http://dx.doi.org/10.1037/0033-295X.115.2.336

Schneider, B., & McDonald, S.-K. (2006). *Scale-up in education: Ideas in principle* (Vol. 1). Lanham, MD: Rowman & Littlefield.

Shapiro, J. A., & Neuberg, S. L. (2007). From stereotype threat to stereotype threats: Implications of a multi-threat framework for causes, moderators, mediators, consequences, and interventions. *Personality & Social Psychology Review, 11,* 107–130. http://dx.doi.org/10.1177/1088868306294790

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology, 38,* 183–242. http://dx.doi.org/10.1016/S0065-2601(06)38004-5

Sherman, D. K., Cohen, G. L., Nelson, L. D., Nussbaum, A. D., Bunyan, D. P., & Garcia, J. (2009). Affirmed yet unaware: Exploring the role of awareness in the process of self-affirmation. *Journal of Personality and Social Psychology, 97,* 745–764. http://dx.doi.org/10.1037/a0015451

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., . . . Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology, 104,* 591–618. http://dx.doi.org/10.1037/a0031495

Shnabel, N., Purdie-Vaughns, V., Cook, J. E., Garcia, J., & Cohen, G. L. (2013). Demystifying values-affirmation interventions: Writing about social belonging is a key to buffering against identity threat. *Personality and Social Psychology Bulletin, 39,* 663–676. http://dx.doi.org/10.1177/0146167213480816

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139175043

Silverman, A., Logel, C., & Cohen, G. L. (2013). Self-affirmation as a deliberate coping strategy: The moderating role of choice. *Journal of Experimental Social Psychology, 49,* 93–98. http://dx.doi.org/10.1016/j.jesp.2012.08.005

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology, 21,* 261–302. http://dx.doi.org/10.1016/S0065-2601(08)60229-4

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629. http://dx.doi.org/10.1037/0003-066X.52.6.613

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811. http://dx.doi.org/10.1037/0022-3514.69.5.797

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances*

in Experimental Social Psychology, 34, 379–440. http://dx.doi.org/10.1016/S0065-2601(02)80009-0

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin, 37,* 1055–1067. http://dx.doi.org/10.1177/0146167211406506

Tibbetts, Y., Harackiewicz, J. M., Canning, E. A., Boston, J. S., Priniski, S. J., & Hyde, J. S. (2016). Affirming independence: Exploring mechanisms underlying a values affirmation intervention for first-generation students. *Journal of Personality and Social Psychology, 110,* 635–659. http://dx.doi.org/10.1037/pspa0000049

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science, 23,* 73–82. http://dx.doi.org/10.1177/0963721413512856

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39,* 456–467. http://dx.doi.org/10.1016/S0022-1031(03)00019-2

Walton, G. M., Paunesku, D., & Dweck, C. S. (2012). Expandable selves. In M. R. Leary & J. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 141–154). New York, NY: Guilford Press.

Weick, K. E. (1976). Educational Organizations as loosely coupled systems. *Administrative Science Quarterly, 21,* 1–19. http://dx.doi.org/10.2307/2391875

Wilson, T. D. (2011). *Redirect: The surprising new science of psychological change.* New York, NY: Little, Brown.

Yeager, D. S., & Walton, G. M. (2011). Social–psychological interventions in education: They're not magic. *Review of Educational Research, 81,* 267–301. http://dx.doi.org/10.3102/0034654311405999

Received November 20, 2015
Revision received April 27, 2016
Accepted May 14, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 3, 425–438

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000144

Long-Term Positive Effects of Repeating a Year in School: Six-Year Longitudinal Study of Self-Beliefs, Anxiety, Social Relations, School Grades, and Test Scores

Herbert W. Marsh
Australian Catholic University and King Saud University

Reinhard Pekrun
University of Munich and Australian Catholic University

Philip D. Parker
Australian Catholic University

Kou Murayama
University of Reading and Kochi University of Technology

Jiesi Guo and Theresa Dicke
Australian Catholic University

Stephanie Lichtenfeld
University of Munich

Consistently with a priori predictions, school retention (repeating a year in school) had largely positive effects for a diverse range of 10 outcomes (e.g., math self-concept, self-efficacy, anxiety, relations with teachers, parents and peers, school grades, and standardized achievement test scores). The design, based on a large, representative sample of German students (N = 1,325, M age = 11.75 years at Year 5) measured each year during the first 5 years of secondary school, was particularly strong. It featured 4 independent retention groups (different groups of students, each repeating 1 of the 4 first years of secondary school; total N = 103), with multiple posttest waves to evaluate short- and long-term effects, controlling for covariates (gender, age, socioeconomic status, primary school grades, IQ) and 1 or more sets of 10 outcomes collected prior to retention. Tests of developmental invariance demonstrated that the effects of retention (controlling for covariates and preretention outcomes) were highly consistent across this potentially volatile early to middle adolescent period; largely positive effects in the first year following retention were maintained in subsequent school years following retention. Particularly considering that these results are contrary to at least some of the accepted wisdom about school retention, the findings have important implications for educational researchers, policymakers, and parents.

Keywords: math self-concept, achievement, grade retention, social comparison

Supplemental materials: http://dx.doi.org/10.1037/edu0000144.supp

Grade retention is the practice of requiring a student in a given grade or year in school to repeat the same grade level in the following year (Allen, Chen, Willson, & Hughes, 2009). Allen et al. (2009) note that the use of retention as an educational intervention, particularly in the United States, has fluctuated since the early 1900s, reaching a peak in the 1970s before declining in the 1980s, and then increasing rapidly in the 1990s—apparently in response to the standards-based reform movement following the publication of *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983). Marsh (2016) also noted that, on the basis of the Programme for International Student Assessment (PISA) data, there is substantial country-to-country variation in the use of retention.

Social Comparison Theory

Marsh (2016) evaluated the effects of de facto retention (starting school late or repeating a grade) on academic self-concept from the perspective of social comparison theory. Theoretical models such as social comparison theory, adaptation level theory, and range-frequency theory (e.g., Huguet et al., 2009; Marsh, 2016; Marsh et

This article was published Online First August 15, 2016.

Herbert W. Marsh, Institute for Positive Psychology and Education, Australian Catholic University and Faculty of Education, King Saud University; Reinhard Pekrun, Department of Psychology, University of Munich and Institute for Positive Psychology and Education, Australian Catholic University; Philip D. Parker, Institute for Positive Psychology and Education, Australian Catholic University; Kou Murayama, Department of Psychology, University of Reading and Research Unit of Psychology, Education & Technology, Kochi University of Technology; Jiesi Guo and Theresa Dicke, Institute for Positive Psychology and Education, Australian Catholic University; Stephanie Lichtenfeld, Department of Psychology, University of Munich.

This research was supported by four grants from the German Research Foundation (DFG) to Reinhard Pekrun (PE 320/11-1, PE 320/11-2, PE 320/11-3, PE 320/11-4). We thank the German Data Processing and Research Center (DPC) of the International Association for the Evaluation of Educational Achievement (IEA) for organizing the sampling and performing the assessments.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Institute for Positive Psychology and Education (IPPE), Australian Catholic University, 25 Barker Street, Strathfield NSW 2135. E-mail: Herb.Marsh@acu.edu.au

al., 2008) posit that students compare their own academic accomplishments with those of their classmates as one basis for academic self-concept formation. Thus, the academic accomplishments of classmates form a frame of reference or standard of comparison that students use to form their own academic self-concepts. Furthermore, there is a growing body of research showing that academic self-concept is reciprocally related to school-based performance measures (e.g., school grades on report cards) in particular, but also to standardized achievement test scores (Guay, Marsh, & Boivin, 2003; Marsh & Craven, 2006), and that academic self-concept might be even more important than achievement in predicting future academic choices (Marsh & Yeung, 1997).

In academic self-concept studies, the frame of reference is typically defined in terms of the academic achievement of classmates. However, for a variety of reasons, such as acceleration or starting school at an early age, students can find themselves in classes with older, more academically advanced students, who might form a more demanding frame of reference than would same-age classmates. Similarly, because of starting school at a later age, or to being held back to repeat a grade, students can find themselves in classes with younger, less academically advanced students than would other students of the same age. In the present investigation, our focus is on the effects of repeating a year in school on a diverse set of self-beliefs, self-perceptions of relations with significant others, school grades, and standardized test scores collected during the first five years of secondary school.

Time to Learn

Although not studied specifically in relation to retention, Bloom (1976) contended that weaker students merely need more time to learn materials than do stronger students, but that once learning is achieved, the differences between more and less able students diminish in terms of subsequent achievement, academic self-beliefs, and motivation to learn. In addition, there is ample evidence that without appropriate intervention, small differences in achievement at any particular stage of education become larger over time, so that the gap between the more and less able students increases. This cumulative disadvantage has reciprocal effects with subsequent motivation, as well as achievement, creating a downward spiral (i.e., the Mathew effect; Stanovich, 1986; Walberg & Tsai, 1983). Hence, we hypothesize that because retained students have an extra year to learn the materials that originally led to their retention, they should be better able to learn those materials in the first year following retention and should also have more positive self-beliefs, giving them a stronger basis for learning new materials and for maintaining positive self-beliefs in subsequent school years.

Grade Retention Effects

Grade Retention Effects on Achievement

Retention effects (i.e., repeating a year in school) have been studied extensively in relation to academic achievement (e.g., Alexander, Entwisle, & Dauber, 2003; Jimerson, 2001; but see Reynolds, 1992; Roderick, 1994; Roderick & Engel, 2001). However, as emphasized by Jimerson and Brown (2013, p. 140), "because of potential short- and long-term effects that grade retention can have on student achievement and socioemotional outcomes, it remains a controversial topic in research and practice." Indeed, there is a general belief, supported by some research evidence, that retention has negative effects on academic achievement (e.g., Hattie, 2012). As emphasized by Allen et al. (2009), this negative view of retention is evident in a policy statement by the National Association of School Psychologists, which "urges schools and parents to seek alternatives to retention that more effectively address the specific instructional needs of academic underachievers" (p. 481).

However, critical design and methodological issues, such as the need for appropriate control groups and controlling for preexisting differences—especially prior achievement, which is inevitably confounded with retention—dictate caution in reaching overarching conclusions such as these (Jimerson & Brown, 2013). Thus, on the basis of their meta-analysis of grade retention studies, in which they controlled for study quality, Allen et al. (2009) reported that their results "challenge the widely held belief that retention has a negative effect on achievement" (p. 480). They found that studies showing negative effects of retention were largely limited to poor quality studies with insufficient control for preexisting differences.

Consistently with the Allen et al. (2009) meta-analysis, a number of publications based on an ongoing longitudinal study challenge the view that retention has negative effects, or else show that negative effects in prior studies are likely the result of inadequate control for selection effects (Cham, Hughes, West, & Im, 2015; Im, Hughes, Kwok, Puckett, & Cerda, 2013; Moser, West, & Hughes, 2012). Using propensity matching to match retained with nonretained (promoted) primary school students, Wu, West, and Hughes (2010) found that retention had short-term positive effects on school-belonging, teacher-rated engagement, and academic self-concept. In a follow-up to this study, Im et al. (2013) found that retained and promoted students, following transition to middle school, did not differ in terms of achievement, engagement, or school-belonging (although they did not report the follow-up measures of academic self-concept considered in the earlier study, a focus of the present investigation). At Year 5, Moser et al. (2012) compared growth trajectories on math and reading achievement for propensity-matched students who had been retained or promoted in Year 1 of primary school. After shifting scores back 1 year to permit same-year-in-school comparisons (what we refer to as "offset" comparisons), the retention group experienced initially higher scores than the nonretained group, assessed on the basis of Year 1 scores. However, the positive retention effects dissipated over time, such that by Year 5, there were no differences between the two groups. The authors also warned that retention effects on achievement might vary, depending on the nature of the measure, and noted that in Year 3, the retained students were more likely to pass a state accountability math test that was closely aligned to the school curriculum (Hughes, Chen, Thoemmes, & Kwok, 2010). Summarizing the results of these multiple publications, 10 years into this longitudinal research program, Cham et al. (2015) concluded that their ongoing research studies "have not supported the popular view within the educational literature that grade retention harms students' educational success. Instead, we have either found advantages for the retained group or have failed to reject the null hypothesis of no difference between the retained and promoted groups" (p. 18).

Cross-National Comparisons

Marsh (2016) recently proposed a frame-of-reference model to evaluate the effects of relative year in school (e.g., being 1 school

year ahead or behind same-age students) based on math constructs and using PISA data from 41 countries. Marsh showed that for countries participating in PISA, students typically are grouped into the same grade or year in school according to their age, rather than to their abilities in general or in particular school subjects. Thus, with the exception of students who start school early or late, those identified as gifted, or those in need of remedial assistance, it is typical for students within the same class to be of a similar age. For example, based on nationally representative samples of 15-year-olds (total $N = 276,165$) from 41 countries (PISA 2003 data), 67% of the students were in their modal year in school for their country (Marsh, 2016). However, for nearly all countries, there were 15-year-old students who were accelerated 1 or more years relative to their modal year in school (e.g., students in Years 11 or 12 when their modal or "age-appropriate" year group was Year 9 or 10), whereas others were in year groups 1 or more years behind their modal year group (e.g., students in Years 7 or 8 when their modal or "age-appropriate" year group was Year 9 or 10). Extending a model of social comparison theory (Marsh et al., 2015; Marsh, Kuyper, Morin, Parker, & Seaton, 2014), Marsh (2016) predicted a priori, and found, that the effects of de facto retention (starting school late or repeating a grade) on math self-concept (MSC) were consistently positive across the 41 countries. These positive effects of de facto retention were reasonably consistent across the 41 countries and individual student characteristics. Relative year in school seemed to be the critical variable. The critical finding for our purposes is that the positive effects on MSC were similar for students who started late or who had been retained previously.

Noting limitations and directions for further research, Marsh (2016) emphasizes that the cross-sectional nature of the PISA data precludes stronger longitudinal models. He argues, however, that for retained students, the uncontrolled, preexisting differences leading to retention would be likely to negatively bias estimates of the positive effects of de facto retention, working against the hypothesized positive effects that he predicted and found. Similarly, the cross-sectional nature of the data precluded longitudinal models that more fully differentiated between de facto retention based on starting school at an older age, and grade retention. Particularly relevant to the present investigation, and from the perspective of educational policy, the reliance on cross-sectional PISA data precluded evaluation of the effects of retention on changes in academic achievement based either on school grades or on standardized test scores.

Rationale for A Priori Research Hypotheses and Research Questions

The German School System and Grade Retention

In Germany, elementary school spans Years 1 to 4, secondary school starts at Year 5, and compulsory schooling ends at Year 9 in most states, including the state of Bavaria, where the present investigation was conducted. There is no tracking in elementary school, but in most states, including Bavaria, students are placed into one of three tracks at the start of secondary school—lower-track schools (*Hauptschule*), medium-track schools (*Realschule*), and higher-track schools (*Gymnasium*)—on the basis of their elementary school achievement. Grade retention is used in elementary school as well as across all secondary school tracks, and is based on students' achievement in main subjects. The number of repeated years per student is limited, and in the present investigation, no students repeated more than one grade. We also note that in the German school system, teachers are very reluctant to use retention in the first 2 years of secondary school. Hence, the majority of retention in our study appeared in Years 7 and 8, rather than Years 5 and 6.

The Structure of the Data

In the present investigation, we evaluate the effects of grade retention (repeating a school year) on a range of psychosocial and achievement outcomes (see Figure 1) for a single cohort of students as they progress through the first 5 years of secondary school. Data was collected from a representative sample of 1,325 students from 42 schools starting the year before the beginning of secondary school: Year 4 school grades in German and math, and then school grades, standardized achievement tests, and psychosocial variables for each of the subsequent 5 years of secondary school (see Figure 1). We evaluated retention in each of four separate groups: those retained at Year 5, a different group of students retained at Year 6, and so forth, noting that no students were retained for more than 1 year (for a discussion of the German school system, tracking, and retention, see Section 1 of the online Supplemental Materials). The study design (see Figure 1) provides a particularly strong foundation for evaluating retention effects on the basis of multiple natural experiments using longitudinal data that provide multiple posttest waves to evaluate short- and long-term effects of retention and multiple pretest waves as controls for all outcomes as well as the covariates (gender, age, socioeconomic status [SES], primary school grades, IQ).

Our main focus is on the four dichotomous grouping variables (see Figure 1) representing those students who repeated a school year in each of the 4 years from Years 5–8. For example, the lagged effects of repeating Year 5 are represented by the path from the grouping variable ("Repeat Year 5" in Figure 1) to outcomes in the immediate subsequent Wave 2 (Lag 1 effects), as well as all effects in the subsequent three waves (Lag 2–4 effects at Waves 3–5; Figure 1). Whereas most students are in Year 6 in Wave 2, the students repeating Year 5 are in Year 5 at Wave 2. It is important to emphasize that there are Lag 1 effects for each of the four retention groups. Thus (see Figure 1), there are separate estimates of Lag 1 effects for students repeating Years 5, 6, 7, and 8 (i.e., the effects of the first year following retention for each of the four retention groups). Similarly, different groups of students repeating Years 6, 7, and 8, have multiple preretention waves of data to control for preexisting differences, and multiple postretention waves to evaluate the short- and long-term effects of retention. This enables us not only to test these Lag 1 effects for each of the four separate groups but also to test the consistency of these lagged effects across the four groups that span this potentially volatile early to middle adolescent period.

An intentionally diverse set of outcomes was considered, including self-belief variables, the focus of the Marsh (2016) study; achievement measures, which have been the focus of most retention studies; anxiety, to represent the emotional response of students to retention; and student self-reports of relations with sig-

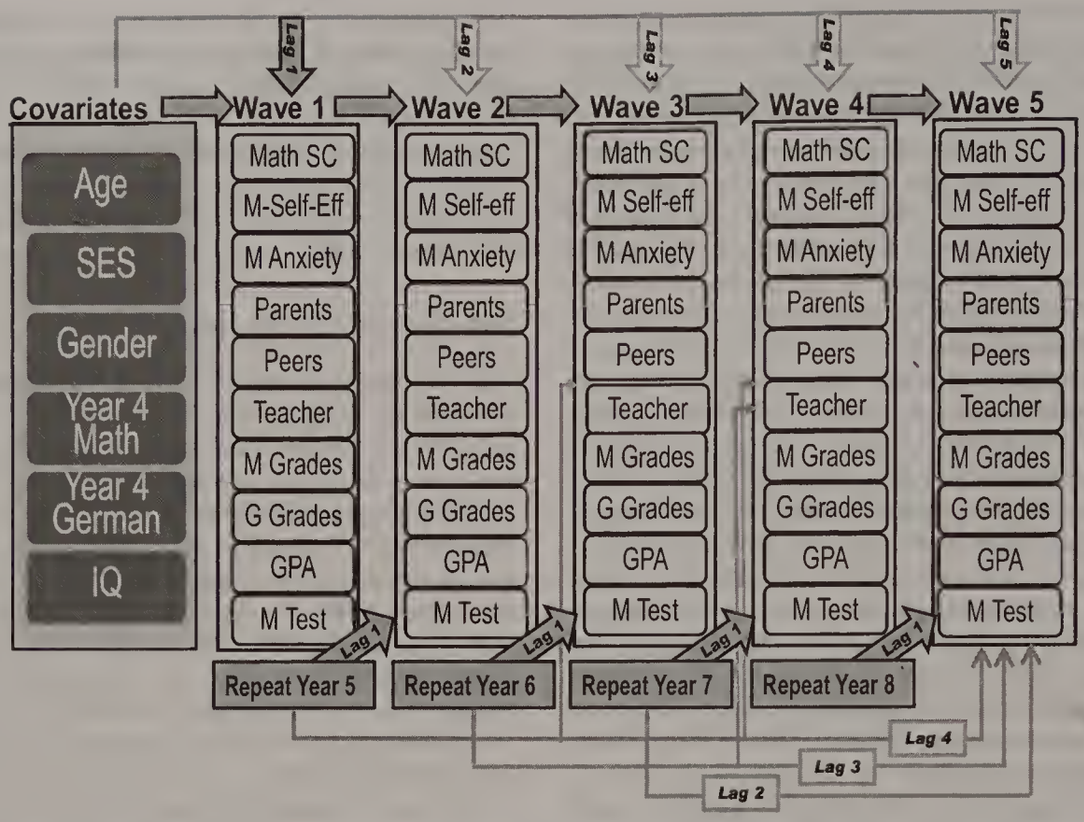


Figure 1. Waves 1 to 5 are the five yearly data collections in this longitudinal study. For students who repeated no grades, the data collections occurred during the first 5 years of secondary school (Years 5 to 9). The same set of 10 outcome variables was collected in each of the five waves. The six covariates are pretest control variables with paths leading from each covariate to all outcomes in Wave 1 (Lag 1 effects, as this is the immediate next wave), Wave 2 (Lag 2 effects), and so forth. Of specific interest are the four dichotomous grouping variables representing students who repeated a school year in each of the four Years 5 to 8. For example, a student repeating Year 5 is tested again in Years 5 (now in Wave 2 rather than Wave 1), 6, 7, and 8 (in Waves 3 to 5). The effect of repeating Year 5 is represented by the path from the grouping variable ("Repeat Year 5") to outcomes in the immediate subsequent wave (Lag 1 effect). The effects of repeating Year 5 are also evaluated in relation to outcomes in Wave 3 (Lag 2 effects, as the outcomes in Wave 3 are two waves following Wave 1), Wave 4 (Lag 3 effects), and Wave 5 (Lag 4 effects). Similarly, different groups of students repeating Years 6 ("Repeat Year 6"), Years 7 ("Repeat Year 7"), and Years 8 ("Repeat Year 8") are each followed in subsequent years to test the effects of repeating grades. For these subsequent groups, Lag 1 effects refer to the effects of repeating a grade on the immediate subsequent wave. For example, for the "Repeat Year 6" group, Lag 1 effects are in relation to outcomes in Wave 3, whereas for the "Repeat Year 7" group, Lag 1 effects are in relation to outcomes in Wave 4. The model depicted is a "full-forward" structural equation model that is saturated, in the sense that all paths are estimated. For example, covariates are predictors of all variables in Waves 1 to 5, Wave 1 variables are predictors of all variables in Waves 2 to 5, and so forth. Within each wave, all variables are correlated. For nonrepeating students, Waves 1 to 5 refer to Years 5 to 9 (the first 5 years of secondary school). Of the 1,325 students considered here, the numbers of students who repeated in each year were: Year 5, n = 10; Year 6, n = 12; Year 7, n = 35; Year 8, n = 45—a total of 103 students, or 7.8% of the total sample of 1,325 students. SES = socioeconomic status; Math SC = self-concept in math; M-Self-Eff = self-efficacy in math; M Anxiety = anxiety in math; Parents = parents work with student in math; Peers = math is valued among peers; Teacher = positive reinforcement from teacher in math; M Grades = final year grade in math; G Grades = final year grade in German; GPA = average grade in other subjects; MTest = standardized math achievement test.

nificant others—parents (academic assistance from parents), teachers (positive teacher support), and peers (peer appreciation of math). (Item wording and reliability estimates, as well as correlations among the multiple factors, are presented in Section 2 of the online Supplemental Materials).

A Developmental Perspective: Developmental Equilibrium Hypothesis

A potentially important limitation of retention research is that it is mostly based on U.S. primary school students, and—even when longitudinal, in terms of following up the effects of retention over multiple school years—typically includes results based on retention in a single school year (see Allen et al., 2009; Holmes & Matthews, 1984; Jimerson, 2001). In this sense, the research lacks a developmental perspective. Here however, we introduce an apparently unique developmental equilibrium perspective, evaluating the consistency of the retention effects over the potentially volatile early to middle adolescent period on the basis of longitudinal data and multiple retention groups. Equilibrium is reached when a system achieves a state of balance between the potentially coun-

terbalancing effects of opposing forces. The application of equilibrium and related terms has a long history in psychological theorizing (Marsh et al., in press). Thus, for example, Marshall, Parker, Ciarrochi, and Heaven (2014) showed that a system of reciprocal effects between self-concept and social support had attained equilibrium by junior high school.

Here we test developmental equilibrium in relation to the invariance of retention effects in each of four separate year groups spanning this early to middle adolescent period. More specifically, we evaluate support for developmental invariance, based on the hypothesis that retention effects are the same for students retained in Years 5, 6, 7, and 8 (see Figure 1). In this sense, our study is longitudinal, in that it covers the entire early to middle adolescent period, but also because it evaluates retention for separate groups of students who had been retained in Years 5, 6, 7, and 8. The German secondary school system starts in year 5, although Years 5 and 6 are often considered part of primary schooling in U.S. studies. Combining the effects of retention across these four groups partly compensates for the typically small sample sizes of retention groups based on retention in a single year, greatly increasing the robustness and statistical power, because of the increased *N* of the results. More importantly, it provides an apparently unique developmental perspective on the question of whether the self-system has achieved a developmental balance in relation to the retention effects, such that they are the same for students retained in Years 5–8.

Research Hypotheses and Questions: Retention Effects in Relation to Specific Outcomes

Math Self-Concept (MSC; Hypotheses 1a and 1b)

Consistently with Marsh (2016), we predict that retention has positive effects on MSC in the first year following grade retention (Lag 1), after controlling for covariates and outcomes from prior waves (Hypothesis 1a). Lag 2–4 effects are the direct effects of retention 2, 3, and 4 years, respectively, following retention, after controlling for Lag 1 effects as well as the effects of covariates and outcomes from the earlier waves. Positive effects at Lags 2–4 would indicate "sleeper effects" (new positive effects, in addition to the positive effects already observed). Nonsignificant effects at Lags 2–4 would indicate that Lag 1 effects were maintained, and negative effects at Lags 2–4 would indicate that Lag 1 effects were not fully maintained. We hypothesize (Hypothesis 1b) that the Lag 2–4 effects of retention will be small and largely nonsignificant—that the initially positive effects of retention on MSC will be maintained.

Self-Efficacy and Anxiety (Hypotheses 2a and 2b)

Although the grounds for these a priori predictions are less clear, both of these variables are strongly related to MSC. On this basis, we anticipate that the effects of retention will be favorable and similar in direction, although perhaps smaller in size, to those predicted for MSC (increased self-efficacy and reduced anxiety) at Lag 1 (Hypothesis 2a), and that these effects will be retained over time (Hypothesis 2b).

Relations With Significant Others (Research Questions 3a and 3b)

Our study includes three variables associated with the positive interactions that students perceive having with significant others (parental assistance, positive teacher support, peer appreciation of math) in relation to math. We leave as research questions the direction of effects of retention on these outcomes at Lag 1 (Research Question 3a) and Lags 2–4 (Research Question 3b), but anticipate that the Lag 1 effects are at least not negative (i.e., are either favorable or are nonsignificant).

School Grades, Lag 1 (Hypothesis 4a, Research Question 4b)

In each year of our study, end-of-year school grades (i.e., school-based performance measures) were collected from school records. For the present purposes, we focus on school grades in math, German (native language), and an average over other subjects. This latter might differ according to the student and year in school (e.g., English, other foreign language, biology, sport, and music). Because retained students study the same materials in the year following retention, Lag 1 retention effects are predicted to be positive and substantial (Hypothesis 4a). An optimistic perspective is that positive Lag 1 effects on school grades are maintained or even increased in subsequent Lags 2–4. However, predicted positive effects at Lag 1 are based on studying the same material for 2 years, whereas Lag 2–4 retention effects are based on students studying new materials for a single year only. Hence, it is entirely possible that the positive effects at Lag 1 will not be fully maintained—that Lag 2–4 retention effects will be negative, offsetting the positive effects at Lag 1, at least in part. Thus, we leave this as a research question, rather than a research hypothesis based on a priori predictions (Research Question 4b).

Standardized Math Test Scores, Same Age Comparisons (Research Questions 5a and b)

In each year of our study, students completed a standardized math test. Although the tests were not specifically based on the school curriculum, in each year, they contained a range of advanced materials suitable to the year in school for nonretained students in each wave of the study. Particularly as retained students have had a chance to learn more fully the materials that they have studied previously, an optimistic perspective would be that Lag 1 retention effects are positive for math test scores. However, because retained students are a year behind their nonretained classmates, they have not studied advanced materials covered in the curriculum that are included in the standardized math test and that have been studied by nonretained students. In this sense, the math test based on same-age comparisons might be considered "unfair" for retained students—at least in terms of inferring what students have learned, relative to the materials that they have actually studied. On the other hand, it could also be argued that the same-age comparisons accurately reflect the fact that repeaters lag behind nonrepeaters in what they have studied. Hence, we leave this as a research question. Particularly given that Lag 1 retention effects on math test scores are left as a research question, there is no basis for predicting Lag 2–4 retention effects; these also are left as a research question.

Offset Math Test Scores, Lag 1 Same-Year-in-School Comparisons (Hypothesis 6a, Research Question 6b)

An alternative perspective on test scores is to compare retained students in each year following retention with nonretained students from the previous wave when they were in the same year in school (see Figure 2). Thus, in this offset strategy (based on comparisons of the same year in school, or what Im et al. [2013, p. 361] refer to as "shifting back" scores), math test scores for retained students repeating Year 5 are compared with test scores from nonrepeaters from the previous wave (when they were also in Year 5) who had studied the same curriculum. Similarly, for each postretention year, for all four retention groups, comparisons based on test scores (but not other outcomes) were "offset" by 1 year, so that comparisons were based on students having completed the same year in school (see Figure 2). For these offset comparisons, we predict that the Lag 1 retention effects will be positive, and more positive than those based on the original (same-age) comparisons (test scores not offset by 1 year; presented in Research Question 5). However, similar to the logic based on school grades (see Research Question 4b), the predicted positive effects for test scores at Lag 1 might not be fully retained over Lags 2–4, and so that we leave this as Research Question 6b.

Method

Sample

Our data are based on the Project for the Analysis of Learning and Achievement in Mathematics (PALMA; Frenzel, Pekrun, Dicke, & Goetz, 2012; Marsh et al., in press; Murayama, Pekrun, Lichtenfeld, & Vom Hofe, 2013; Murayama, Pekrun, Suzuki, Marsh, & Lichtenfeld, 2016; Pekrun et al., 2007; Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, in press), a large-scale longitudinal study investigating the development of math achievement and its determinants during secondary school in Germany. The study was conducted in the German federal state of Bavaria. The present investigation included five measurement waves spanning Years 5 to 9, in addition to school grades from the last year of primary school (Year 4). Data (1,325 students from 42 schools; 50% girls; mean age = 11.75 years at Wave 1, SD = 0.7) were

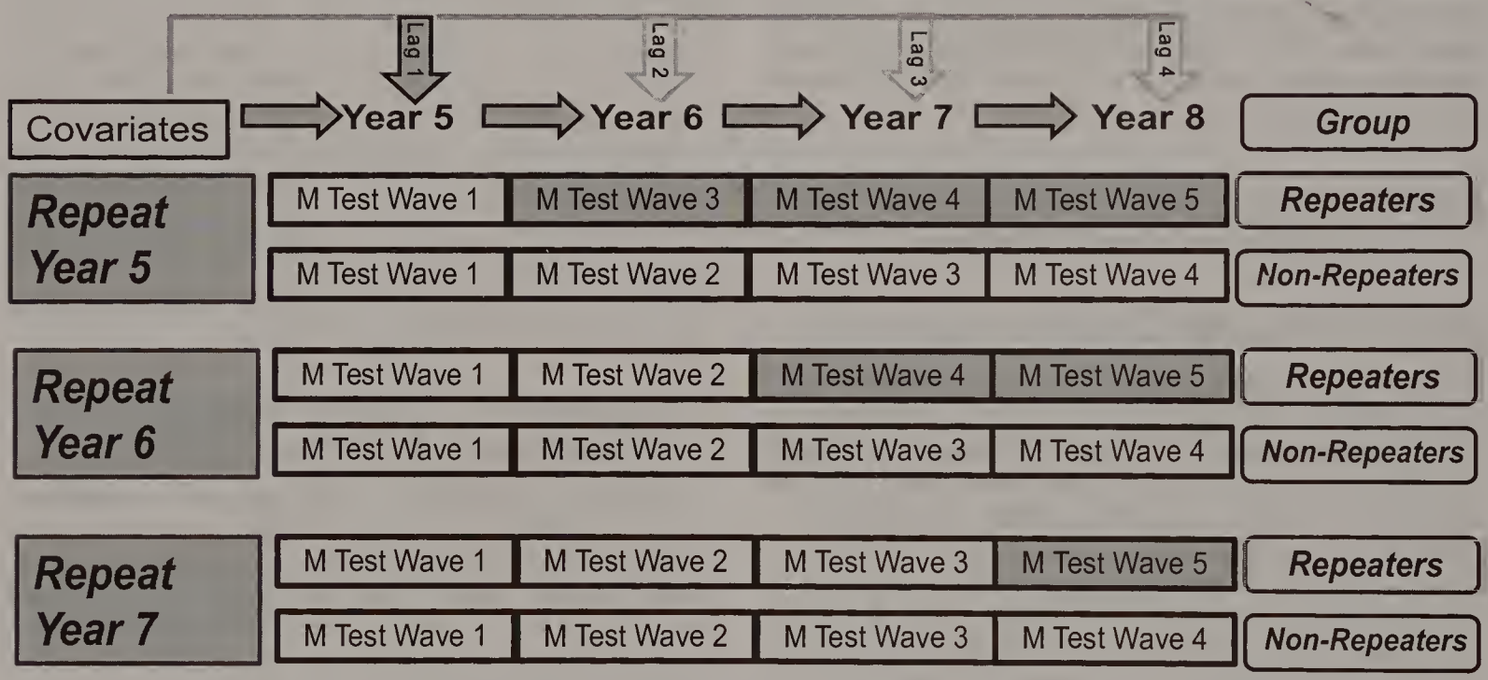


Figure 2. Offset comparisons for standardized math tests (M Tests) in Waves 1 to 5. Depicted is an alternative perspective on test scores in which retained students in each year following retention are compared with nonretained students from the previous wave. For example, math test scores for students repeating Year 5 in Wave 2 were compared with test scores of nonrepeating students when they also completed Year 5 (but in Wave 1 rather than Wave 2). Likewise, Year 6 (Wave 2) math test scores for nonrepeating students are compared with test scores from repeaters who have also just completed Year 6 (but in Wave 3 rather than Wave 2). In this way, math tests are based on the performances of students who have studied the same curriculum. Similarly, for each postretention year (those shaded in gray for the repeater groups) for all four retention groups, comparisons based on test scores (but not other outcomes) were "offset" by 1 year, so that comparisons were based on students having completed the same year in school. Separate analyses were done for each retention group, except for the "repeat Year 8" retention Group, in which this offset strategy was not possible (i.e., there are no Year 9 scores for the retention group that can be compared with the Year 9 scores for the nonrepeater group). In other respects, the offset analysis is like the "full-forward" structural equation model depicted in Figure 1, in that all the same covariates and outcomes are included (only the math test scores are "offset"); all covariates are predictors of all variables in Years 5 to 9, Year 5 variables are predictors of all variables in Years 6 to 9, and so forth. Again, the main focus of the present investigation is on the dichotomous grouping variables representing students who repeated a school year in one of the four Years 5 to 8.

collected from the year before the start of secondary school (Year 4 school grades in German and math), and school grades, standardized achievement tests, and psychosocial variables for each of the subsequent 5 years of secondary school (see Figure 1).

Sampling and assessments were conducted by the Data Processing and Research Center of the International Association for the Evaluation of Educational Achievement. The samples represented the typical student population in the state of Bavaria in terms of student characteristics such as gender, urban versus rural location, and SES (for details, see Pekrun et al., 2007). Students answered the questionnaire toward the end of each successive school year. All instruments were administered in the students' classrooms by trained external test administrators. Participation in the study was voluntary, parental consent was obtained for all students, and the acceptance rate was very high at 91.8%. Surveys were depersonalized to ensure participant confidentiality.

Our central focus is on evaluating the effects of grade retention in each of the first 4 years of secondary school. Because grade retention is not a frequent occurrence, the numbers repeating are relatively small. Of the 1,325 students considered here who participated in all five waves of the study, present investigation the numbers of students who repeated in each year were as follows: Year 5 (n = 10); Year 6 (n = 12); Year 7 (n = 35); Year 8 (n = 45)—a total of 103 students, or 7.8% of the sample. The 103 repeating students did not differ significantly (all $ps > .05$) from the 1,222 nonrepeating students on gender (42% vs. 51% female); school type (43% Gymnasium, 23% Realschule, 23% Hauptschule vs. 40%, 30%, and 29%, respectively); age (11.7 vs. 11.8 years); or family SES (.01 vs. −.02).

In supplemental analyses, we evaluated potential biases associated with missing data after controlling for background variables (see "covariates" in Figure 1) and school type for the 10 outcomes in Year 5. More specifically, we evaluated the main effect of being included in the sample ("include" in online supplemental Table 2; the difference between the 1,325 students in the final sample vs. the 745 students excluded because of missing data); main effect of repeat ("repeat" in online supplemental Table 2; the differences in outcomes for the repeating students compared with those who did not repeat Year 5); and the Repeat × Include Interaction ("IncldxRepeat" in online supplemental Table 2). This last parameter was of particular interest, as it explored whether the difference between repeating and nonrepeating students depended upon whether the students were included in the final sample. The effects of "include" were statistically significant for two of 10 outcomes; those students in the final sample had significantly higher math grades ($p < .01$) and German grades ($p < .05$) than students excluded because of missing data, but did not differ significantly in terms of school grades in other subjects, standardized test scores, or any of the other outcomes. Students had missing data over this 5-year span because of absences on the day of the data collection, and also because families moved. However, we note that there are very strong controls for biases associated with these outcomes, as each of the 10 outcomes was measured in each of the five waves of data. More importantly for present purposes, differences between repeating and continuing students did not depend upon whether the students were or were not included in the final sample. More specifically, differences between the repeating and nonrepeating students on the 10 outcomes in Year 5 did not vary significantly as a function of missing data, thereby supporting the appropriateness of the analyses (see Section 1 of the online Supplemental Materials).

Measures

See Section 2 of the online Supplemental Materials for more detail on measures.

Six psychosocial constructs. At each measurement wave, the same set of items was used to assess MSC, math self-efficacy, math anxiety (Achievement Emotions Questionnaire-Mathematics; see Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011), and student perceptions of significant others—parents (Parental Assistance), teachers (Positive Teacher Support), and peers (Peer Appreciation of Math). All of these multi-item scales were based on self-report responses from students, using a 5-point-Likert scale: *not true at all, hardly true, somewhat true, mostly true, or completely true*. Across the five waves and the six multi-item scales, the 30 coefficient alpha estimates of reliability were generally high (αs varying from .75 to .92; median α = .87) and were consistent over the multiple waves. For ease of interpretation, anxiety scores were reverse scored, so that—consistently with other constructs—higher scores reflect more favorable outcomes. (Item wording and reliability estimates, as well as correlations among the multiple factors, are presented in Section 2 of the online Supplemental Materials).

Math achievement. Students' achievement was measured both in terms of school grades (from Year 4, the last year of primary school, and in Years 5–9, the first 5 years of secondary school) and standardized achievement test scores in math (Years 5–9). School grades were end-of-year final grades obtained from school records. Standardized math achievement was assessed by the PALMA Mathematical Achievement Test (vom Hofe, Kleine, Blum, & Pekrun, 2005). Using both multiple-choice and open-ended items, this test measures students' modeling and algorithmic competencies in arithmetic, algebra, and geometry. In each successive year, the test covered the same content areas, but the number and difficulty of the items increased in line with the year in school completed by nonrepeating students; the number of items increased from 60 to 90 items across the five waves. The obtained achievement scores were scaled using one-parameter logistic item response theory (Rasch scaling; Wu, Adams, Wilson, & Haldane, 2007), and standardized in relation to Year 5 results (i.e., the first measurement point) to establish a common metric across the five waves.

Covariates. Students' school grades in math and German at the end of primary school (Year 4), gender, IQ, age, and SES served as covariates for the overall study. Student IQ was measured using the 25-item nonverbal reasoning subtest of the German adaptation of Thorndike's Cognitive Abilities Test (Heller & Perleth, 2000). SES was assessed by parent report using the Erikson Goldthorpe Portocarero social class scheme (Erikson, Goldthorpe, & Portocarero, 1979), which consists of ordered categories of parental occupational status; higher values represent higher social class.

Statistical Analyses

All analyses were done with Mplus 7.3 (Muthén & Muthén, 2008–2014, Version 7). We used the robust maximum likelihood

estimator, which is robust against violations of normality assumptions. All analyses were based on manifest variables, using the complex design option to account for nesting of students within schools. As is typical in large longitudinal field studies, some students had missing data for at least one of the measurement waves, due primarily to absence or to changing schools. Because of the nature of the data analyses (particularly the "offset" comparison of math test scores), analyses were based on the 1,325 students who participated in all five waves. For this group, the relatively small amounts of missing data (less than 1% for each variable) were handled with full information maximum likelihood, the default option in Mplus.

The primary analysis was a "full-forward" structural equation model that is saturated, in the sense that all paths are estimated (see Figure 1). For example, covariates are predictors of all variables in Years 5–9, Year 5 variables are predictors of all variables in Years 6–9, and so forth. Within each wave, all variables were correlated. A specific focus is the four dichotomous grouping variables representing students who repeated a school year in one of the 4 years from Years 5–8. For example, a student repeating Year 5 is tested again in Year 5 (now in Wave 2 rather than Wave 1), and again in Years 6, 7, and 8 (in Waves 3–5). The effect of repeating Year 5 is represented by the path from the grouping variable ("Repeat Year 5") to outcomes in the immediate subsequent wave (Lag 1 effects), as well as all subsequent waves (effects at Lags 2–4). Similarly, different groups of students, repeating Years 6, 7, and 8, are each followed up in subsequent years, to test the effects of retention.

In order to facilitate interpretation of the results, all covariates and Year 5 outcomes were standardized ($M = 0$, $SD = 1$) across the entire sample. Outcomes for Years 6–9 were then standardized in relation to mean values of each construct in Year 5, so that measurement in relation to a common metric was retained. The four grouping variables representing retention were scored as 1 = retention and 0 = nonretention. Hence, the unstandardized coefficients associated with each of these variables represent the difference between the two groups in relation to Year 5 standard deviation units, after controlling for covariates and outcomes in all waves prior to retention for each of the retention groups—hereafter referred to as effect sizes (ESs)—scaled so that higher scores reflect more favorable outcomes. As noted earlier (see discussion of research questions, and Hypotheses 6 and 7), retention effects on standardized achievement tests were evaluated in relation to both same-age comparisons (e.g., comparing results of retained Year 5 students with those of nonretained Year 6 students who are of a similar age) and same-year-in-school comparisons (e.g., comparing results of retained Year 5 students with nonretained students when they also were in Year 5; see Figure 2).

Preliminary Analyses: Evaluation of Developmental Invariance Hypothesis

The path model depicted in Figure 1 is a "full forward" structural equation model that is completely saturated, with degrees of freedom equal to zero; all paths relating variables in different waves are estimated, as are all correlations and correlated residuals relating variables within each wave. We evaluated two alternative models to summarize the retention effects. In the "means model" we used the model constraint option in Mplus to compute the mean ES across the relevant retention groups for each outcome, along with the standard error and a test as to whether the mean was significantly different from zero. Thus, for example, the mean ES for MSC was the mean retention effect averaged across the four retention groups (i.e., students retained in Years 5, 6, 7, and 8). Importantly, this model is still saturated, in that it did not impose any constraints. However, it provides a much stronger, more robust test of the overall retention effects, in that the test of the mean across retention groups is based on a larger N than tests of each group separately, compensating in part for the small number of retained students in each retention group.

In order to more formally evaluate the invariance of retention effects, we next tested a "developmental invariance" model in which all lagged effects were constrained to be the same across the four retention groups. Thus, for example, Lag 1 retention effects for MSC were constrained to be the same for the different groups of students who had been retained in Years 5, 6, 7, and 8, respectively. This highly constrained, parsimonious model imposed a total of 60 invariance constraints. Particularly given the large number of constraints, the fit of this model was remarkably good, providing strong support for the developmental invariance of retention effects across the four retention groups. Not surprisingly, the mean ESs (based on the means model) and the invariant ESs (based on the developmental invariance model) were similar, and both provided a parsimonious summary of the retention effects. For the present purposes, we focus on results based on the statistically stronger developmental invariance model, but results for the means model—including the estimates for each of the year groups considered separately, as well as details about the fit of the developmental invariance—are presented in the online Supplemental Materials (Section 4).

Results

Effects of Retention

Math self-concept (Hypotheses 1a and 1b). Consistently with Hypothesis 1a, the effects of retention on MSC in the first year following retention (invariant Lag 1 effects) were positive and statistically significant (ES = .597, Table 1). Lag 2–4 effects reflect the direct effect of the intervention after controlling for outcomes from all previous waves, including the Lag 1 effects; positive effects reflect "sleeper" effects, negative effects reflect a significant diminishing of the positive effects at Lag 1, and nonsignificant effects reflect maintenance of the positive effects at Lag 1. Consistently with Hypothesis 1b, the *ES*s for Lags 2–4 were nonsignificant (maintenance of Lag 1 effects).

Self-efficacy and anxiety (Hypotheses 2a and 2b). Consistently with Hypothesis 2a, the effects of retention on these outcomes were significantly positive (noting that anxiety was reverse scored so that higher values reflect less anxiety). However, ESs (.359 for self-efficacy, .293 for anxiety; Table 1) were smaller than for MSC. Consistently with Hypothesis 2b, Lag 2–4 ESs were nonsignificant for both self-efficacy (maintenance of Lag 1 effects), although for anxiety effects there was a positive Lag 4 effect (a

Table 1
Short-Term (Lag 1) and Long-Term (Lags 2–4) Effects of Grade Retention Across 4 Years of Secondary School

10 outcomes	Invariant Lag 1 effects (ESs)		Invariant Lag 2 effects (ESs)		Invariant Lag 3 effects (ESs)		Invariant Lag 4 effects (ESs)	
	Effect size	*SE*	Effect size	*SE*	Effect size	*SE*	Effect size	*SE*
Math self-concept	**.597****	.094	.148	.116	−.113	.215	.405	.210
Math self-efficacy	**.359****	.084	.079	.122	−.155	.161	.128	.326
Math anxiety	**.293****	.092	.207	.117	−.100	.159	**.656****	.217
Parents	.173	.110	.008	.129	.277	.236	.336	.180
Peer	.023	.094	−.020	.154	.002	.203	.365	.270
Teacher	**.305****	.099	.149	.133	−.007	.166	.209	.194
Math grades	**1.010****	.119	−.033	.134	.077	.240	.396	.210
German grades	**.454****	.068	−.059	.117	−.025	.160	.191	.203
Grade Point Average	**.452****	.054	−.092	.080	.053	.110	−.187	.181
Math test	**−.188***	.076	−.143	.100	.059	.091	.222	.178
Total	**.348****	.042	.024	.059	.027	.075	**.272****	.090

Note. Analysis based on Figure 1 (where variables are defined), a "full-forward" structural equation model that is saturated, in the sense that all paths are estimated and correlations within each wave are estimates. Based on support of developmental invariance model, effect sizes (ESs) were constrained to be invariant over the four retention groups. ESs are the "direct effects" of repeating a grade on each outcome variable, controlling for covariates and all outcomes from prior waves. Lag 1 paths are those for the first year after repeating a grade; Lag 2 paths are the effects on the second year following grade retention, controlling for outcomes from all prior waves—including Lag 1 effects, and so forth. All outcome variables are standardized in relation to Year 5 (Wave 1) values. ESs that are statistically significant ($p < .05$) in relation to their standard errors (*SE*s) are in bold.
$^{*}p < .05$. $^{**}p < .01$.

positive sleeper effect), even though Lag 2 and 3 effects were nonsignificant.

Relations with significant others (Research Questions 3a and 3b). Lag 1 ESs for the effects of student perceptions of positive teacher support were significantly positive (ES = .305), whereas the nonsignificant Lags 2–4 effects indicated that these positive effects of retention were maintained in subsequent school years. There were no statistically significant effects (Lags 1–4) of retention for perceptions of parental assistance or peer appreciation of math.

School grades (Hypothesis 4a and Research Question 4b). Retention effects were evaluated for end-of-year school grades for math and for German (required subjects), and an average grade over other subjects (grade point average [GPA]). Lag 1 retention effects were significantly positive for all three measures of school grades (ESs = .452 to 1.010). The results were particularly large for math school grades (mean ES = 1.010), reflecting stronger controls for preexisting differences in math, because of the focus of the study on math (i.e., other outcomes, including test scores, were math-specific). Although we anticipated that the corresponding Lag 2–4 effects might be negative (but left this as a research question), these effects were all nonsignificant, demonstrating that the substantial positive effects of retention on school grades in the first year following retention were maintained in subsequent school years.

Standardized math tests, same-age comparisons (Research Questions 5a and b). Retention effects were evaluated in relation to standardized achievement test scores collected in each year of the study. We anticipated that these Lag 1 effects based on same age comparisons might inappropriately disadvantage retained students (who had not studied some of the advanced materials covered by nonretained students), but left this as a research question. Indeed, Lag 1 effects for math test scores were significantly negative (ES = −.188), although the size of the effect was much smaller than the corresponding positive effect on school grades (ES = +1.010). Lag 2–4 effects for test scores were nonsignificant, indicating that the small negative effects of retention on test scores were maintained (see Table 1).

Standardized math tests, same-year-in-school comparisons (Hypothesis 6a and Research Question 6b). In an alternative perspective on test scores (see Figure 2 and Table 2), we compared test scores of retained students in each year following retention with those of nonretained students in the previous wave (i.e., same-year-in-school comparisons). Thus, test scores for the retained groups were compared with those in nonretained groups who had completed the same year in school and studied the same curriculum, but on the basis of data from one wave earlier. Because of the nature of the offset comparisons (see Table 1), these had to be conducted separately for retention groups in Years 5–7 (and were not possible for the "repeat Year 8" retention group; see discussion in Table 2). Consistently with Hypothesis 6a (see Table 2), Lag 1 ESs were more positive for these offset comparisons (based on the same year in school) than were those based on the same wave (same-age comparisons, evaluated in Research Question 5a). For these offset comparisons, all six ESs (based on total effects in Table 2) were positive (.053 to .677; $M = .341$) in favor of the retention group, and three were statistically significant. In summary, when test scores for retained students were compared with those of other students in the same year group, there were significantly positive effects of retention.

Summary of Results

Given the persistent belief that retention has negative effects, the most important finding here is that in research based on a particularly strong and more appropriate design, the effects of retention were mostly positive, and almost none were significantly negative. Indeed, for the critical Lag 1 effects based on the first year following the intervention, only one of the 10 effects was significantly negative ($.05 < p < .01$), and seven were significantly

Table 2
Comparison of Effects of Repeating a Year in School Based on the Original Math Tests (Same-Age Comparisons) and Math Tests Offset by 1 Year (Same-Year-in-School Comparisons)

		Time (number of waves following retention)					
		Total effects			Direct effects		
Repeating group	Comparison	Lag 1	Lag 2	Lag 3	Lag 1	Lag 2	Lag 3
Repeat Year 5	Original	−.078 (.206)	−.076 (.175)	.034 (.149)	−.078 (.206)	−.152 (.102)	−.107 (.189)
	Offset	.101 (.110)	**.603** (.146)	.242 (.219)	.101 (.110)	**.442** (.146)	.024 (.156)
Repeat Year 6	Original	.022 (.143)	−.079 (.148)		.022 (.143)	−.193 (.152)	
	Offset	**.677** (.155)	**.371** (.151)		**.677** (.155)	−.022 (.157)	
Repeat Year 7	Original	**−.253** (.106)			**−.253** (.106)		
	Offset	.053 (.165)			.053 (.165)		

Note. The analyses presented here are based on Figure 1 (where variables are defined) and on the analyses in Table 1, but differ in several important aspects. First, separate analyses were done for each of the four groups of repeaters. Second, as with the analyses in Table 1, outcomes following the repeated year are controlled for covariates and outcomes from all previous waves, and correlations within each wave are estimated. Most importantly, math standardized test scores (but none of the other outcomes) for repeating groups were offset by one wave, such that repeating students were compared with nonrepeating students who had completed the same year in school (see Figure 2). Thus, for students who repeated Year 5, math test scores for Waves 3–5 (when they were in Years 6–8) were compared with math test scores for nonrepeating students for Waves 2–4 (when they were also in Years 6–8). For each of the repeating groups, separate analyses are presented for the original math test scores and for 1-year offset math test scores. Results are presented both for the total effect of retention (controlling for covariates and outcomes prior to retention) and for direct effects (controlling for covariates, outcomes prior to retention, and outcomes following retention, as in Table 1). Results involving Wave 5 are not presented because the offset transformation for retention groups uses Wave 5 math test scores as Wave 4 (see Figure 2). Standard errors of each path are presented in parentheses, and statistically significant paths, $p < .05$, are presented in bold.

positive ($p < .01$). Averaged across the 10 outcomes, the mean of Lag 1 effects was statistically significant (.384). Evaluation of Lag 2–4 effects of retention demonstrate that these Lag 1 effects were maintained, or in the case of anxiety, improved further in subsequent years. Although our focus has been on the invariant estimates across the four retention groups, it is also relevant to look at the results for each of the four groups separately (see Section 4 of the online Supplemental Materials). For the critical 40 Lag 1 effects (i.e., four retention groups × 10 outcomes) based on the first year following the intervention, only one of the 40 effects was significantly negative ($.05 < p < .01$). Furthermore, none of the mean effects for any of the 10 outcomes averaged across the four retention groups were significantly negative. In contrast, 23 of 40 effects were significantly positive; the mean effects averaged across the four groups were significantly positive for 6 of 10 outcomes, as was the grand mean effect averaged across all outcomes (.384).

Consistently with Marsh (2016), the effects of retention on MSC were positive (*M* Lag 1 ES = .597), and the results were generally favorable for self-efficacy and anxiety. However, perhaps surprisingly, the results were even more positive for math school grades (*M* Lag 1 ES = 1.010); the retention effects were also positive for other school grade measures. Retention effects for relations with significant others were positive, but only student perceptions of teacher support were statistically significant.

Discussion, Limitations, and Directions for Further Research

Developmental Equilibrium

The developmental perspective adopted here is apparently new in retention research and has important implications. Consistently with the developmental equilibrium hypothesis, the largely positive effects of retention, and the maintenance of these effects, were highly consistent across different groups of students who had been retained in Years 5, 6, 7, and 8. Support for this hypothesis not only supports the robustness and consistency of the positive retention effects but also indicates that the self-system has achieved equilibrium in relation to retention effects over this potentially volatile period. Because this is an apparently new strategy in retention research, it is important that future research tests the generalizability of these retention effects and extends to students of other ages.

Retention Effects for School Grades

The substantial Lag 1 effects in favor of retained students, particularly for math grades (*M* ES = 1.010) require further consideration. These Lag 1 effects might be argued to advantage the retained students unfairly, because they had studied the same curriculum for 2 consecutive years. However, this would not be the case for effects in subsequent years following retention (i.e., Lags 2–4). Hence, because of the finding that Lag 2–4 effects for math grades were nonsignificant, the initial positive Lag 1 effects were maintained in subsequent school years. The positive retention effects were larger for math school grades than for school grades in German, and the GPA based on other school subjects. However, this difference can be explained, at least in part, by the focus of this study on math, with the consequence that there were stronger controls for preexisting differences in relation to math than there were for other school subjects—particularly those included in the GPA measure, for which controls in relation to some school subjects were limited. As noted earlier, residual preexisting differences are likely to advantage nonrepeating students; this potential bias was apparently larger for nonmath outcomes.

Retention Effects for Standardized Math Tests—Same Age Versus Same Year (Offset) Comparisons

Retention effects for math standardized test scores were the least positive, and were slightly negative when based on same-age comparisons (–.188; Table 1). However, these results apparently reflected—at least in part—an apparent unfairness in these comparisons, in the sense that retained students were being tested on advanced materials that they had not covered in their studies, whereas these materials had been covered by nonretained students. In an alternative strategy, we argued that retained student results should be compared with those of students who had completed the same year in school—what we refer to as offset (or same-year-in-school) comparisons. Thus, for example, results for the Year 5 retention group were compared with the results of students who had completed Year 5 in the previous wave, rather than with the results for these same students after they had completed Year 6. For these offset comparisons, the total effects for the retention group were all positive (M ES = .341)—significantly so for three of six comparisons.

Interpretation of these results on the basis of standardized test scores is not straightforward. On the one hand, it might be argued that the same-age comparisons unfairly favored nonretained students, as they were taught materials covered in the test that had not been taught to the retained students. Furthermore, this same issue was present in all subsequent years (i.e., retained students were always 1 year behind the nonretained students). However, the standardized math test in our study focused on generic skills appropriate for the age groups, and was not specifically based on the school curriculum. This is similar to the rationale for PISA tests. Hence, the advantage for nonretained students in our study is likely to be much smaller than in studies that use tests specifically based on the curriculum covered by the nonretained students.

On the other hand, it might be argued that our offset comparisons unfairly advantage the retained students, who have been taught the same materials for 2 consecutive years. Again, this potential advantage would likely be even larger for a test that more closely reflected the curriculum—in this case, for the class completed by the retained students, rather than the nonretained students. However, even to the extent that such comparisons advantaged the retained students, this advantage would only be relevant for Lag 1 comparisons: In subsequent school years, previously retained students would only have been taught the new materials in a single school year. Hence, it is important to emphasize that for the offset comparisons, our results show that the positive effects of retention in the first year following retention (Lag 1 results) were maintained over subsequent school years (Lags 2–4). Furthermore, even the offset comparisons have a potential bias in favor of the nonretained students, in that the comparison group for evaluating retention (i.e., the nonretained students) is truncated, excluding all the poorest performing students who were originally part of that cohort (i.e., the retained students). Hence, the offset comparisons provide important evidence for the benefits of retention, even for standardized test scores.

The offset approach used here, to test for the effects of retention on the basis of standardized test scores, is not the only strategy to circumvent potentially biased comparisons in favor of nonretained students. For example, an alternative approach might be to compare the results of retained students with those of their new classmates following retention (that is, those who, while in the same year in school, are typically 1 year younger), rather than their former classmates, prior to retention. This approach would have the advantage of comparing retained students with a whole cohort of new students, rather than with a truncated cohort that excluded retained students, but would have the disadvantage that controlling for preexisting differences might be more problematic. Although there is apparently no completely satisfactory solution to this problem, it is critical that future research provide reasonable controls in relation to potentially biased comparisons of retained and nonretained students in respect of materials that have only been taught to nonretained students. Similarly, systematic reviews and meta-analyses of the effects of retention need to distinguish results on the basis of how this issue is addressed in primary studies (see Allen et al., 2009).

Potential Process Mechanisms to Explain Positive Retention Effects

Although they are beyond the scope of the present investigation, it is important to explore process mechanisms to explain the positive retention effects: These can be the basis of further research. The Marsh (2016) study, which was a starting point for the present investigation, used frame of reference models (e.g., social comparison theory; Marsh et al., 2015; Marsh et al., 2014) to predict positive effects of retention (and negative effects of acceleration) on academic self-concept. In this respect, the present investigation is consistent with previous findings. Furthermore, there is a growing body of research demonstrating that academic self-concept and achievement—particularly school grades, but also test scores—are reciprocally related (e.g., Marsh & Craven, 2006; Pinxten, Marsh, De Fraine, Van Den Noortgate, & Van Damme, 2014). Relatedly, the fact that students do so much better, in terms of school grades, after repeating a year in school, is likely to reinforce their MSC and psychological adjustment more generally. Hence, this theoretical rationale explains the results of the present investigation—at least in part.

Although apparently there have been no retention studies focusing mainly on the time required to master new materials, or on Matthew effects, these theoretical perspectives appear to be relevant. There is clear theoretical and empirical evidence from mastery learning interventions that weaker students might merely need more time to master new material, material that can be mastered more quickly by stronger students (Carroll, 1989; Kulik, Kulik, & Bangert-Drowns, 1990). There is also theoretical and empirical research on the Matthew effect showing that without intervention, students who fall behind at any particular stage in schooling tend to fall behind even further in subsequent school years (e.g., Stanovich, 1986; Walberg & Tsai, 1983). According to Bloom (1976), if weak students are given sufficient time and resources to achieve mastery, the differences between more and less able students will diminish, and achieving mastery has potentially profound effects on positive self-beliefs and motivations to learn. Similarly, Stanovich (1986) argued that early intervention is critical to break the vicious cycle created by Matthew effects. Consistent with these theoretical and empirical perspectives, the fact that retained students had an extra year to learn the materials that had led to their retention not only helped them to learn those materials more effectively in the first year following retention but

also resulted in more positive self-beliefs and gave them a stronger basis for learning new materials in subsequent school years. Hence, retention can be seen as a potentially useful intervention to counter the negative consequences of failure to learn critical academic materials.

We also note that retained students tend to be more mature (i.e., a year older than their new classmates following retention). Indeed, it is curious that there seems to be widespread support for holding students back when they start school so that they are among the oldest in their class, rather than the youngest (also referred to as "academic red shirting"; see Gladwell, 2008), but the opposite view prevails in terms of holding students back by repeating a school year when they have not adequately mastered the materials (the so-called "old for grade" hypothesis; see Im et al., 2013). However, Marsh (2016) argues that the advantage of being relatively older than classmates in terms of academic self-concept is similar for students who started late and those who repeat a year in school, and that this pattern of results has broad cross-national generalizability. Our results are consistent with those conclusions, but extend them in important new directions—particularly in relation to academic achievement and the long-term maintenance of short-term benefits of retention.

Limitations

A major limitation of the present investigation is the relatively small number of retained students, particularly for any given school year. Although this limitation is inherent in the nature of this research, it means that very large samples are needed to obtain even modest numbers of retained students. To some extent, our design compensated for this limitation by considering multiple retention groups. Relatedly, although the longitudinal design is clearly stronger than cross-sectional comparisons and comparisons based on just two waves of data for a single retention group, causal interpretations of correlational data should always be made cautiously. As noted by Allen et al. (2009), the most critical problem in making causal inferences about grade retention is the absence of randomized control trials that control for preretention differences, although they also note that "for obvious reasons, random assignment of students to the 'treatments' of retention and promotion is neither feasible nor ethical" (p. 481). Nevertheless, our design was particularly powerful in that we controlled for a strong set of covariates and a complete set of outcome variables for up to three waves of preretention data, and evaluated postretention results for the same set of outcomes for up to 3 years following retention. Furthermore, uncontrolled preexisting differences between retained and nonretained students were likely to favor nonretained students, thus working against our a priori hypotheses and supporting results in favor of retention. Importantly, the results were consistent across multiple groups who had been retained in Years 5–8; this is consistent with our developmental equilibrium hypothesis.

Our study was based on students at the start of secondary school from a single German state, so there is clearly a need to replicate the results in different settings and with different age groups. We also note as a potential limitation the large number of students with missing data for at least one of the five waves of this longitudinal study. However, we do note that at least the positive effects of retention on academic self-concept results replicate and extend the results of Marsh (2016), which showed that the positive effects of retention generalize reasonably well across nationally representative samples of 15-year-olds from 41 different countries.

As emphasized by Reardon (2011), Parker, Jerrim, Schoon, and Marsh (2016), and many others, there is clear evidence of a steadily increasing gap between academically advantaged and disadvantaged students, particularly in the United States but also in many other industrialized countries as well. There is also evidence (Micklewright & Schnepf, 2007) that the median achievement levels of countries as a whole are negatively related to the gap between the advantaged and disadvantaged. Hence, countries all over the world are trying to devise policies to decrease the gap. From this perspective, the strategic use of retention might be an effective strategy to counter this trend. However, we also note that there is an economic component of costs to the school system associated with retention and providing an extra year of schooling. There is also perhaps a "cost" to individual students in terms of potentially delaying their entry into the labor market. Hence, although this is obviously beyond the scope of our study, cost–benefit analyses would be needed to evaluate whether the costs are outweighed by the benefits.

Summary and Implications

Our results have important implications for educational researchers, but also for parents, teachers, and educational policymakers. Indeed, schools in different countries, and even in different geographic regions of the same country, use diverse strategies in relation to school retention, apparently without fully understanding the implications of these policy practices in relation to a variety of psychosocial variables and academic achievement measures such as those considered here, which have long-term implications for academic choice and accomplishments. Particularly because the results of the present investigation are contrary to at least some accepted wisdom in relation to retention, as understood by parents and schools, there is a need for further research to more fully evaluate the generalizability and construct validity of the interpretations offered here. However, our results clearly refute any simplistic conclusion that retention is necessarily "bad."

References

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (2003). *On the success of failure: A reassessment of the effects of retention in the primary school grades*. New York, NY: Cambridge University Press.

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multi-level analysis. *Educational Evaluation and Policy Analysis, 31,* 480–499. http://dx.doi.org/10.3102/0162373709352239

Bloom, B. S. (1976). *Human characteristics and school learning*. New York, NY: McGraw-Hill.

Carroll, J. B. (1989). The Carroll Model: A 25-year retrospective and prospective view. *Educational Researcher, 18,* 26–31. http://dx.doi.org/10.3102/0013189X018001026

Cham, H., Hughes, J. N., West, S. G., & Im, M. H. (2015). Effect of retention in elementary grades on grade 9 motivation for educational attainment. *Journal of School Psychology, 53,* 7–24. http://dx.doi.org/10.1016/j.jsp.2014.10.001

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in 3 western European societies: England, France and Sweden. *The British Journal of Sociology, 30,* 415–441. http://dx.doi.org/10.2307/589632

Frenzel, A. C., Pekrun, R., Dicke, A. L., & Goetz, T. (2012). Beyond quantitative decline: Conceptual shifts in adolescents' development of interest in mathematics. *Developmental Psychology, 48,* 1069–1082. http://dx.doi.org/10.1037/a0026895

Gladwell, M. (2008). *Outliers*. New York, NY: Little, Brown.

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95,* 124–136. http://dx.doi.org/10.1037/0022-0663.95.1.124

Hattie, J. A. (2012). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Abingdon, UK: Routledge.

Heller, K. A., & Perleth, C. (2000). *Kognitiver fähigkeitstest für 4.-12. Klassen, Revision (KFT 4–12 + R)* [Cognitive ability test, revised version (KFT 4–12 + R)]. Göttingen, Germany: Hogrefe.

Holmes, C. T., & Matthews, K. M. (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research, 54,* 225–236. http://dx.doi.org/10.3102/00346543054002225

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Hughes, J. N., Chen, Q., Thoemmes, F., & Kwok, O. M. (2010). An investigation of the relationship between retention in first grade and performance on high stakes tests in third grade. *Educational Evaluation and Policy Analysis, 32,* 166–182. http://dx.doi.org/10.3102/0162373710367682

Huguet, P., Dumas, F., Marsh, H., Wheeler, L., Seaton, M., Nezlek, J., . . . Régner, I. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology, 97,* 156–170. http://dx.doi.org/10.1037/a0015558

Im, M. H., Hughes, J. N., Kwok, O. M., Puckett, S., & Cerda, C. A. (2013). Effect of retention in elementary grades on transition to middle school. *Journal of School Psychology, 51,* 349–365. http://dx.doi.org/10.1016/j.jsp.2013.01.004

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review, 30,* 420–437.

Jimerson, S. R., & Brown, J. A. (2013). Grade retention. In J. A. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 42–44). New York, NY: Routledge.

Kulik, C. L., Kulik, J. A., & Bangert-Drowns, J. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60,* 265–299. http://dx.doi.org/10.3102/00346543060002265

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31,* 357–365. http://dx.doi.org/10.1177/0165025407077757

Marsh, H. W. (2016). Cross-cultural generalizability of year in school effects: Negative effects of acceleration and positive effects of retention on academic self-concept. *Journal of Educational Psychology, 108,* 256–273. http://dx.doi.org/10.1037/edu0000059

Marsh, H. W., Abduljabbar, A. S., Morin, A. J. S., Parker, P., Abdelfattah, F., Nagengast, B., & Abu-Hilal, M. M. (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology, 107,* 258–271. http://dx.doi.org/10.1037/a0037485

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Hillsdale, NJ: Erlbaum.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163. http://dx.doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald* (pp. 275–340). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Kuyper, H., Morin, A. J. S., Parker, P. D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50–66. http://dx.doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Pekrun, R., Lichtenfeld, S., Guo, J., Arens, A. K., & Murayama, K. (in press). Breaking the double-edged sword of effort/trying hard: Developmental equilibrium and longitudinal relations among effort, achievement, and academic self-concept. *Developmental Psychology*.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 319–350. http://dx.doi.org/10.1007/s10648-008-9075-6

Marsh, H. W., & Yeung, A. S. (1997). Coursework selection: The effects of academic self-concept and achievement. *American Educational Research Journal, 34,* 691–720. http://dx.doi.org/10.3102/00028312034004691

Marshall, S. L., Parker, P. D., Ciarrochi, J., & Heaven, P. C. L. (2014). Is self-esteem a cause or consequence of social support? A 4-year longitudinal study. *Child Development, 85,* 1275–1291. http://dx.doi.org/10.1111/cdev.12176

Micklewright, J., & Schnepf, S. (2007). Inequalities in industrialised countries. In S. P. Jenkins & J. Micklewright (Eds.), *Inequality and poverty re-examined* (pp. 129–145). Oxford, UK: Oxford University Press.

Moser, S. E., West, S. G., & Hughes, J. N. (2012). Trajectories of math and reading achievement in low achieving children in elementary school: Effects of early and later retention in grade. *Journal of Educational Psychology, 104,* 603–621. http://dx.doi.org/10.1037/a0027571

Murayama, K., Pekrun, R., Lichtenfeld, S., & Vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development, 84,* 1475–1490. http://dx.doi.org/10.1111/cdev.12036

Murayama, K., Pekrun, R., Suzuki, M., Marsh, H., & Lichtenfeld, S. (2016). Don't aim too high for your kids: Parental over-aspiration undermines students' learning in mathematics. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspp0000079

Muthén, L. K., & Muthén, B. (2008–2014). *Mplus user's guide*. Los Angeles CA: Author.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U. S. Government Printing Office.

Parker, P. D., Jerrim, J., Schoon, I., & Marsh, H. W. (2016). A multination study of socioeconomic inequality in expectations for progression to higher education: The role of between-school tracking and ability stratification. *American Educational Research Journal*. Advance online publication.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36,* 36–48. http://dx.doi.org/10.1016/j.cedpsych.2010.10.002

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (in press). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development.*

Pekrun, R., vom Hofe, R., Blum, W., Frenzel, A. C., Goetz, T., & Wartha, S. (2007). Development of mathematical competencies in adolescence: The PALMA longitudinal study. In M. Prenzel (Ed.), *Studies on the educational quality of schools* (pp. 17–37). Münster, Germany: Waxmann.

Pinxten, M., Marsh, H. W., De Fraine, B., Van Den Noortgate, W., & Van Damme, J. (2014). Enjoying mathematics or feeling competent in mathematics? Reciprocal effects on mathematics achievement and perceived math effort expenditure. *British Journal of Educational Psychology, 84,* 152–174. http://dx.doi.org/10.1111/bjep.12028

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In R. Murnane & G. Duncan (Eds.), *Whither opportunity? Rising inequality and the uncertain life chances of low-income children* (pp. 91–116). New York, NY: Russell Sage Foundation Press.

Reynolds, A. J. (1992). Grade retention and school adjustment: An explanatory analysis. *Educational Evaluation and Policy Analysis, 14,* 101–121. http://dx.doi.org/10.3102/01623737014002101

Roderick, M. (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal, 31,* 729–759. http://dx.doi.org/10.3102/00028312031004729

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes testing. *Educational Evaluation and Policy Analysis, 23,* 197–227. http://dx.doi.org/10.3102/01623737023003197

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407. http://dx.doi.org/10.1598/RRQ.21.4.1

vom Hofe, R., Kleine, M., Blum, W., & Pekrun, R. (2005). On the role of "Grundvorstellungen" for the development of mathematical literacy. First results of the longitudinal study PALMA. *Mediterranean Journal for Research in Mathematics Education, 4,* 67–84.

vom Hofe, R., Pekrun, R., Kleine, M., & Götz, T. (2002). Projekt zur analyse der leistungsentwicklung in mathematik (PALMA): Konstruktion des Regensburger mathematikleistungstests für 5.-10. Klassen [Project for the analysis of learning and achievement in mathematics (PALMA): Development of the Regensburg mathematics achievement test for grades 5 to 10]. *Zeitschrift fur Padagogik, 45*(Beiheft), 83–100.

Walberg, H. J., & Tsai, S. (1983). Matthew effects in education. *American Educational Research Journal, 20,* 359–373.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modeling software* [Computer software]. Retrieved from Australian Council for Educational Research website https://www.acer.edu.au/conquest

Wu, W., West, S. G., & Hughes, J. N. (2010). Effect of grade retention in first grade on psychosocial outcomes. *Journal of Educational Psychology, 102,* 135–152. http://dx.doi.org/10.1037/a0016664

Received January 19, 2016
Revision received May 12, 2016
Accepted May 19, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 3, 439–449

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000140

Academic Competencies: Their Interrelatedness and Gender Differences at Their High End

Sebastian Bergold, Heike Wendt, Daniel Kasper, and Ricarda Steinmayr
Technical University Dortmund

The present study investigated (a) how a latent profile analysis based on representative data of N = 74,868 4th graders from 17 European countries would cluster the students on the basis of their reading, mathematics, and science achievement test scores; (b) whether there would be gender differences at various competency levels, especially among the top performers; (c) and whether societal gender equity might account for possible cross-national variation in the gender ratios among the top performers. The latent profile analysis revealed an international model with 7 profiles. Across these profiles, the test scores of all achievement domains progressively and consistently increased. Thus, consistent with our expectations, (a) the profiles differed only in their individuals' overall performance level across all academic competencies and not in their individuals' performance profile shape. From the national samples, the vast majority of the students could be reliably assigned to 1 of the profiles of the international model. Inspection of the gender ratios revealed (b) that boys were overrepresented at both ends of the competency spectrum. However, there was (c) some cross-national variation in the gender ratios among the top performers, which could be partly explained by women's access to education and labor market participation. The interrelatedness of academic competencies and its practical implications, the role of gender equity as a possible cause of gender differences among the top performers, and directions for future research are discussed.

Keywords: academic achievement, gender differences, TIMSS and PIRLS, gender equity, latent profile analysis

Both in educational research and practice and by students themselves, the belief prevails that most students exhibit considerable strengths in some domains (e.g., reading) and at the same time considerable weaknesses in other domains (e.g., mathematics; Marsh & Hau, 2004; Wang, Eccles, & Kenny, 2013). In stark contrast to this belief stand empirical findings documenting that competencies in different domains such as reading, mathematics, and science are highly intercorrelated (e.g., Reilly, 2012; Rindermann, 2007). These findings imply that higher competencies in one domain are likely to be accompanied by higher competencies in the other domains. Studies that investigate student profiles across different academic competency domains are lacking, even though studies using an intraindividual approach provide a more comprehensive picture of students' overall academic competency than studies only focusing on one or two competencies (e.g., Brunner et al., 2013). Due to the high intercorrelations among different competencies, we hypothesized that students' competency profiles would only differ in the absolute values across competency domains but not in the shape of different competencies relative to each other when investigating reading, mathematics, and science competencies. Academic competencies in different domains are partly shaped by the numerous determinants located at the country, the school, the classroom, and the student level (e.g., Byrnes & Miller, 2007). As these processes seem to work equally in comparable schooling systems, we expected that the students' profile patterns could be replicated between countries with comparable schooling systems.

Moreover, we examined gender differences in the competency profiles. We were especially interested in the profile representing those students having the highest competencies across all domains. Previous research has found that boys were overrepresented in the upper tail of the mathematics and science distribution but girls in reading competencies (e.g., Hedges & Nowell, 1995; Nowell & Hedges, 1998). Hitherto, the question has been unanswered whether more boys or girls are present in the right tail of the ability distribution when considering all three competencies simultaneously. Answering this question is important for the ongoing debate on gender differences in academic competencies and on women's underrepresentation in scientific careers (Ceci, Williams, & Barnett, 2009; Hyde, 2014). Although research suggests that gender differences in payment and promotion opportunities as well as women's interests, career preferences, and variety of choice options considerably contribute to women's underrepresentation in science (Ceci et al., 2009; Ferriman, Lubinski, & Benbow, 2009; Hunt, 2016; Wang et al., 2013), high competencies in academic domains such as reading, mathematics, and science (and spatial ability; e.g., Wai, Lubinski, & Benbow, 2009) are a further im-

This article was published Online First July 18, 2016.

Sebastian Bergold, Department of Psychology, Technical University Dortmund; Heike Wendt and Daniel Kasper, Institute for School Development Research, Technical University Dortmund; Ricarda Steinmayr, Department of Psychology, Technical University Dortmund.

Correspondence concerning this article should be addressed to Sebastian Bergold, Department of Psychology, Technical University Dortmund, Emil-Figge-Straße 50, D-44227 Dortmund, Germany. E-mail: sebastian.bergold@tu-dortmund.de

portant prerequisite for scientific careers (Ceci et al., 2009). Furthermore, it is important to investigate whether the gender differences in the highest competency profile are related to any cultural specifics of the countries we considered. Doing so serves to shed further light on the reasons for why males or females are overrepresented at the highest end of mathematics, reading, and science performance.

In the present study, we used representative data of elementary school children from the Trends in International Mathematics and Science Study (TIMSS) 2011 and the Progress in International Reading Literacy Study (PIRLS) 2011 from 17 European countries that were members of the European Union in 2011. Both studies were run simultaneously and the same children were investigated with the competency tests from both TIMSS and PIRLS. Applying a person-centered approach, we investigated (a) whether a latent profile analysis (LPA) on the basis of mathematics, science, and reading achievement test scores would reveal profiles differing in their individuals' overall performance level rather than in their individuals' performance profile shape; (b) whether there are gender differences in any of the profiles and especially in the profile representing the highest performance level; and (c) whether we would find cross-national variability of the gender ratios in the highest competency profile and whether this variability might be explained by societal gender equity.

Toward a Comprehensive Perspective on Academic Competencies

Different studies find academic achievement tests to be highly intercorrelated. For example, the reading, mathematics, and science scores from PISA were found to intercorrelate from $r = .75$ to $r = .81$ (Reilly, 2012) or even from $r = .95$ to $r = .99$ (Rindermann, 2007). In TIMSS and PIRLS 2011, the correlations were lower but still substantial, ranging from $r = .54$ to $r = .74$ (Bos et al., 2012). Students doing better than other students in, for example, reading do on average also better in mathematics and science. This is because mathematics and science tasks might to a certain degree require reading skills or because reading, mathematics, and science might all be influenced by superordinate factors.

Indeed, children's academic competencies are the results of a myriad of determinants located at the country, the school, the classroom, and the student level (see, e.g., Byrnes & Miller, 2007). A lot of these shape interindividual differences in students' academic achievement in a rather uniform way, that is, in about the same manner across different domains. This might explain why countries achieving certain levels in one domain tend to achieve similar levels also in other domains (e.g., Martin, Mullis, Foy, & Stanco, 2012; Mullis, Martin, Foy, & Arora, 2012; Mullis, Martin, Foy, & Drucker, 2012; Organization for Economic Cooperation and Development [OECD], 2014). Even though these country-level differences are not yet well understood, some explanations are discussed that refer to factors influencing the entire educational system, for example, differences in the states' investments in education (e.g., OECD, 2014). Furthermore, variables located at the school and at the classroom level (see Chung, 2015; Kellaghan, 2015) or—highly correlated with indices of school quality—students' family background variables, such as the socioeconomic status (SES), are correlated with performance in different domains (see Kurtz-Costes, 2015; OECD, 2014; Sirin, 2005). The same is true for student-level determinants. At the student level, the most important single determinant of academic achievement is cognitive ability, which is highly correlated with competencies in all domains (e.g., Deary, Strand, Smith, & Fernandes, 2007; Jensen, 1998). Cognitive ability is not only associated with school performance but also with variables influencing school success at different levels such as family background or school environment (see, e.g., Chung, 2015). In addition, motivational variables and school-relevant personality variables such as conscientiousness, openness for experience, or test anxiety are also important predictors of academic achievement across domains and are correlated with other determinants on the same and other levels (e.g., Hembree, 1988; Steinmayr, Dinger, & Spinath, 2010, 2012; Steinmayr & Spinath, 2009).

Summing up, there are several, partly interrelated, variables having a general influence on academic competencies across domains that might explain the high correlations between competencies in different domains. As they are found across countries with comparable schooling systems, we assume that they impact on students' competencies all in the same way. More precisely, when investigating a large sample comprising students from a wide range of different settings—homes, classes, schools, and countries—one would expect that students doing well in one domain should also tend to do well in other domains. Students should have rather balanced ability profiles instead of unbalanced profile shapes, and this should hold across the countries.

For the investigation of ability profiles, a person-centered approach should be the method of choice. However, to our knowledge, so far only one study has used such an approach. Wang et al. (2013) conducted a LPA with SAT reading and mathematics scores from 1,490 12th graders. They identified five profiles: high math/high verbal scores ($n = 298$), high math/moderate verbal scores ($n = 373$), moderate math/moderate verbal scores ($n = 402$), low math/high verbal scores ($n = 298$), and low math/low verbal scores ($n = 119$). At first glance, these results seem partly to contradict the expectation of balanced ability profiles. However, Wang et al.'s (2013) sample was preselected for ability level, in that it was composed of "intellectually able, college-bound U.S. students" (p. 771). In samples of preselected and more abled individuals, the influence of important cross-domain determinants such as family background or school quality might vanish. Furthermore, because of the preselection, Wang et al.'s (2013) terms *high*, *moderate*, and *low scores* have to be seen in relation to the ability group they investigated. For example, a value of 655 points (on the SAT subtest scale from 200 to 800 points with an average of 500 points) was termed *moderate* (p. 772). However, when compared with scores that a population-representative sample would obtain, 655 points would reflect an above-average result, just like the scores termed *high*. Thus, when bearing the full ability range in mind, the discrepancies between the domains were not as large as the titles of the profiles might have suggested. Moreover, in high ability ranges, narrow abilities often gain in importance relatively to broad abilities (e.g., Reynolds, Keith, & Beretvas, 2010). In addition, the reliability of the measurement often shrinks as the ability level becomes more extreme, especially at the subtest level. This leads to more heterogeneous ability profiles than would be found in a population-representative sample (Rost, 2013). In our study, we relied on representative and large samples from

different countries, comprising students from a wide range of homes, classrooms, schools, and countries. Furthermore, we also considered competencies in science in addition to competencies in reading and mathematics.

Gender Differences on Academic Performance Tests

Academic competencies—in particular, the core competencies reading, mathematics, and science—predict a variety of important life outcomes, such as subsequent school grades, participation and academic success in higher education, and success on the labor market (e.g., Kuncel, Hezlett, & Ones, 2001; Richardson, Abraham, & Bond, 2012). Thus, the question whether there are gender differences at different levels of academic competencies is of high practical relevance. This is particularly true for gender differences at the high end of academic competencies because these are likely to contribute to the explanation of why there are far more men than women among the top performers in key societal positions such as in research (Ceci et al., 2009).

A substantial body of research based on large representative samples and a number of meta-analyses have already investigated gender differences on mathematics, science, and reading achievement tests. Some of them have addressed gender differences in mean scores only. However, gender ratios in the tails of the ability distribution can be due to both mean and variance differences between genders (we assume normally distributed scores for each gender; e.g., Feingold, 1992; Hedges & Friedman, 1993). Thus, mean differences, variance differences, and resulting gender ratios in the upper and in the lower tail as well as in the middle range of mathematics, science, and reading ability distributions, respectively, will be reviewed in the following.

Analyses of gender differences in mean mathematics test scores have revealed at most a small effect in favor of boys (e.g., Else-Quest, Hyde, & Linn, 2010; Guiso, Monte, Sapienza, & Zingales, 2008; Hedges & Nowell, 1995; Reilly, Neumann, & Andrews, 2015; but see Brunner, Krauss, & Kunter, 2008). Inspection of the variance ratios (i.e., male variance divided by female variance) showed that boys displayed higher variance on math scores than girls (e.g., Lindberg, Hyde, Petersen, & Linn, 2010; Reilly et al., 2015). Variance ratios (VRs) were only small in most cases and typically ranged from 1.05 to 1.20, that is, boys were 5% to 20% more variable than girls. However, greater male variance resulted in a considerable overrepresentation of boys in the upper tail of the mathematics score distribution, all the more if it was combined with a small mean difference in favor of boys. For example, Nowell and Hedges (1998) found that a d of 0.09 and a VR of 1.13 resulted in a gender ratio of 1.62:1 in favor of boys at the top 5%, of 2:1 at the top 3%, and of 2.62:1 at the top 1% of math achievers. Because of boys' slightly greater mean, both in the lower tail and in the middle range of the distribution, there were either no gender differences or girls were slightly overrepresented (Hedges & Nowell, 1995; Machin & Pekkarinen, 2008; Nowell & Hedges, 1998). Similar results were also found for science (e.g., Hedges & Nowell, 1995; Nowell & Hedges, 1998; Reilly et al., 2015).

Whereas studies revealed a substantial overrepresentation of boys among high mathematics and science achievers, they showed the opposite pattern for high reading achievers. For instance, Nowell and Hedges (1998) found girls to outscore boys on mean reading achievement by $d = 0.23$ in 1992. Although boys again tended to have a greater variability than girls (VR = 1.14), the considerable mean difference resulted in a female overrepresentation in the upper tail of the reading ability distribution (see also Machin & Pekkarinen, 2008; Reilly, 2012). Because of boys' greater variability, however, the gender ratio in favor of girls did not increase but decreased with a stricter cut-off criterion (top 5%: 1.33:1; top 3%: 1.27:1; top 1%: 1.11:1). In the lower tail of the distribution, boys were clearly overrepresented (lowest 10%: 1.85; lowest 5%: 1.97), whereas they were slightly underrepresented in the middle range (see also Hedges & Friedman, 1993; Machin & Pekkarinen, 2008).

Thus, gender differences vary depending on the ability level and domains investigated. No study so far has examined gender differences among top achievers considering more than one domain. Doing so is important because all of the academic core competencies are predictors of future success. With the present study, we seek to close this research gap, and investigate whether and, if so, to which extent boys are overrepresented among the highest achieving students when reading competencies (in which girls prevail among the highest achievers) were considered in addition to mathematics and science competencies.

What Causes Gender Differences on Academic Performance Tests?

The reasons for gender differences in academic achievement are not yet well understood. Several explanations have been suggested, which can roughly be subdivided into the biological and the sociocultural account. The biological account encompasses evolutionary, genetic, hormonal, and brain-related explanations. However, evidence for biological theories is mixed and studies with sufficient numbers of highly abled individuals are missing (see, e.g., Ceci et al., 2009; Hyde, 2014, for an overview).

Sociocultural theory posits that gender differences are driven by social influences such as societal gender equity (e.g., equity in labor division, women's access to education; see Ceci et al., 2009; Hyde, 2014). Although the results reviewed above have indicated overall greater means and variability for boys in mathematics and science as well as greater means for girls in reading, studies with large representative international samples have also found considerable variability in gender differences across countries (as well as across ethnicities and cohorts; e.g., Else-Quest et al., 2010; Guiso et al., 2008; Penner, 2008; Reilly, 2012). Consequently, cross-national variability in both mean and variance differences resulted in cross-national variability in gender ratios in the upper tail of the ability distributions, which are especially important with regard to gender differences among the later top performers in society. For example, Guiso et al. (2008) showed that, for every boy at the top 5% mathematics achievers, there were, on average, 0.6 girls. However, this ratio ranged from 0.4 (Korea) to 1.1 (Indonesia; see also Penner, 2008). This variability in findings suggests that the sociocultural context is a crucial factor related to gender differences in achievement test scores. In most cases, gender differences in mathematics and science tended to be smaller, and gender differences in reading tended to be greater, in countries with higher gender equity (Else-Quest et al., 2010; Guiso et al., 2008; Penner, 2008; Reilly, 2012).

However, the studies reported above have examined each academic competency separately. Yet, all three academic competencies are predictors of important life outcomes and are substantially intercorrelated. Therefore, in the evaluation of gender differences, they should be considered simultaneously. We are not aware of any study that pursued this target applying a person-centered approach.

Hypotheses and Research Questions

Evidence from research on academic achievement suggests that different academic competencies are substantially interrelated. Thus, we expected (a) that, in a sample of students from a wide range of settings, we would find student profiles that differ in their overall performance level across all three domains mathematics, science, and reading rather than in the shape of their performance profile. Because the intercorrelations among academic domains are observed internationally (Brunner et al., 2013; Reilly, 2012; Rindermann, 2007); we expected (b) the profile solution to be valid also for the individual countries. We also examined (c) the gender ratios in the different profiles, especially whether boys would be overrepresented among the top performers. To gain insight into the mechanisms underlying the genesis of unequal gender ratios at the high end of academic competencies, we examined (d) whether the gender ratios in the highest profile would vary across the countries. If this would be the case, then we expected (e) that this variation could at least partly be explained by societal gender equity.

Method

Sample

We used representative data from the TIMSS 2011 and the PIRLS 2011. Assessments for TIMSS and PIRLS are usually conducted independently of each other, executed in different cycles and focusing on different academic competencies. In 2011, however, TIMSS and PIRLS coincided with each other for the first time and were thus in some countries conducted mutually, collecting mathematics, science, and reading test data from the same students.

Overall, 32 countries participated in the combined TIMSS 2011 and PIRLS 2011 assessments with nationally representative samples of fourth graders. To foster Europe's economic growth, in 2010 the European Commission launched Europe 2020, an economic program that also entailed the Education and Training 2020 program. Within this program, a common strategic educational framework was set up for all members of the European Union to support them in further developing their educational systems to promote lifelong learning and achieve higher school completion and employment rates as well as greater educational justice (European Commission, 2015). Thus, the member states of the European Union follow the same broad educational framework and have relatively comparable educational systems, at least with regard to elementary school. We aimed to take advantage of this fact because, if cross-national differences were observable, these could not be explained by differences between educational systems. This provides ideal prerequisites to examine the role of societal gender equity in the genesis of gender differences. Therefore, we used the data from the $k = 17$ countries that both took part in the combined TIMSS/PIRLS 2011 assessments and were members of the European Union when data collection took place (i.e., Austria, Czech Republic, Finland, Germany, Hungary, Ireland, Italy, Lithuania, Malta, Northern Ireland, Poland, Portugal, Romania, Slovak Republic, Slovenia, Spain, and Sweden). The 17 national samples resulted in a total sample of $N = 74{,}868$ fourth-grade students (36,655 girls and 38,213 boys; see Table 1) from 2,704 schools. Due to the sampling procedure implemented in TIMSS and PIRLS, it was ensured that the children in the different countries were all

Table 1
Students' N, Mean Age, and Mean Scores (Standard Errors) in Reading, Mathematics, and Science

	N				Reading		Mathematics		Science	
Country	Girls	Boys	Total	Mean age	Girls	Boys	Girls	Boys	Girls	Boys
Austria	2,232	2,355	4,587	10.3	533 (2.2)*	525 (2.7)	504 (2.7)	513 (3.3)*	525 (2.8)	538 (3.6)*
Czech Republic	2,159	2,274	4,433	10.3	549 (2.5)*	542 (2.5)	505 (2.8)	516 (2.7)*	529 (2.9)	544 (2.7)*
Finland	2,223	2,318	4,541	10.8	578 (2.3)*	558 (2.2)	542 (2.5)	549 (2.9)*	570 (2.9)	570 (3.0)
Germany	1,940	1,988	3,928	10.4	545 (2.3)*	537 (2.7)	523 (2.7)	532 (2.6)*	522 (3.0)	534 (3.2)*
Hungary	2,533	2,616	5,149	10.6	547 (3.2)*	532 (3.2)	514 (3.6)	517 (3.9)	532 (4.0)	537 (3.9)
Ireland	2,165	2,218	4,383	10.3	559 (2.9)*	544 (3.0)	526 (3.7)	529 (3.3)	516 (4.0)	516 (4.6)
Italy	2,067	2,058	4,125	9.7	543 (2.4)	540 (2.7)	503 (3.1)	512 (2.9)*	520 (3.2)	528 (3.0)*
Lithuania	2,200	2,384	4,584	10.7	537 (2.4)*	520 (2.4)	533 (2.6)	534 (2.9)	514 (2.4)	515 (3.0)
Malta	1,694	1,798	3,492	9.8	486 (2.4)*	468 (2.0)	492 (1.6)	499 (2.1)*	443 (2.2)	449 (2.8)*
Northern Ireland	1,717	1,752	3,469	10.4	567 (2.5)*	550 (3.2)	562 (3.3)	563 (3.6)	517 (3.2)	516 (3.2)
Poland	2,394	2,568	4,962	9.9	533 (2.5)*	519 (2.7)	476 (2.4)	486 (2.5)*	502 (3.0)	508 (2.9)*
Portugal	1,957	2,034	3,991	10.0	548 (3.0)*	534 (2.8)	529 (4.1)	535 (3.4)	519 (4.6)	524 (3.8)
Romania	2,246	2,397	4,643	10.8	510 (4.8)*	495 (4.3)	481 (6.7)	484 (5.9)	505 (6.9)	506 (5.7)
Slovak Republic	2,736	2,825	5,561	10.4	540 (3.1)*	530 (2.8)	503 (4.0)	511 (3.9)*	528 (4.3)	536 (3.6)*
Slovenia	2,115	2,318	4,433	9.8	539 (2.2)*	523 (2.7)	508 (2.2)	518 (3.1)*	517 (2.8)	523 (3.4)
Spain	2,021	2,084	4,105	9.8	516 (2.5)	511 (2.8)	477 (3.1)	488 (3.4)*	500 (2.8)	510 (3.7)*
Sweden	2,193	2,289	4,482	10.7	549 (2.4)*	535 (2.5)	501 (2.5)	506 (2.4)	532 (3.0)	535 (3.2)
Total	36,592	38,276	74,868	10.3	540 (.6)*	528 (.6)	516 (.7)	522 (.7)*	518 (.8)	524 (.8)*

Note. Achievement test data after weighting to obtain nationally representative samples. Achievement test scores ranged from 5.0 to 870.2.
* Statistically significantly ($p < .05$) higher than the values of the other gender.

comparable on their age and on their amount of schooling. All countries applied a strict sampling procedure to allow analyses of nationally representative data. Sample selection was strictly monitored in every country to preserve high quality sampling standards (for more details regarding the sampling procedure, see Martin & Mullis, 2012).

Measures

Academic competencies. In TIMSS and PIRLS 2011, the students were administered academic competency measures assessing students' proficiency in reading (12 reading passages, 146 items), mathematics (180 items), and science (206 items).[1] The achievement tests had been developed on the basis of advice from substantive and statistical expert panels and the results of extensive pilot studies according to internationally aligned frameworks, ensuring, inter alia, measurement invariance across genders. Whereas the reading test items assessed reading literacy, both the mathematics and the science test items were constructed on the basis of a core curriculum that was comparable for all countries (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009).

The test items were administered in a multimatrix design in which 14 different booklets covering both mathematics and science and 13 booklets covering reading were randomly assigned to the students. Each student worked on two reading passages and approximately 24 to 30 mathematics and 24 to 30 science items. Although there was some time restriction, the TIMSS and PIRLS tests were primarily designed as power tests (Mullis, Martin, Kennedy, et al., 2009; Mullis, Martin, Ruddock, et al., 2009).

Gender equity. We used domain-specific gender equity indicators that would be theoretically expected to explain cross-national differences in academic achievement (Else-Quest & Grabe, 2012). These indicators were (a) women's access to education, as measured by the ratio of women and men with at least a secondary education level (typically at least 9 years of education completed) and by the ratio of the net tertiary school enrollment rates of women and of men (tertiary education programs comprised either theory-based programs to qualify for academic high-skill professions, with a duration of at least 3 years, or somewhat more practical programs to qualify for a direct entry into the labor force, with a duration of at least 2 years). Women's access to education "reflects the value of girls' education" in a society (Else-Quest & Grabe, 2012, p. 137). This value could lead girls and women to accordingly value their own academic development and consequently to engage in or to disengage from it. The second indicator was (b) women's participation in the labor market (as measured by the ratio of women's and men's participation in the labor market in the working age group between 15 and 64; participation regardless of how many hours worked). The last indicator of gender equity was (c) women's share of research positions (in percent of all research positions; head count, comprising both part-time and full-time employment). The two latter indicators might explain gender differences in academic achievement because working women, especially when working in research, might serve as role models for girls. This could make girls feel more self-confident about their abilities and their opportunities to participate later in society as an equal member of it, which could in turn promote their academic achievement (Else-Quest & Grabe, 2012; Else-Quest et al., 2010). Thereby, women's share of research positions, that is, in an academic top position, might be especially significant for the gender ratios at the top academic competency level.

The data for these indicators were taken from the United Nations Educational, Scientific, and Cultural Organization Institute for Statistics (http://data.uis.unesco.org); from the Organization for Economic Cooperation and Development (http://stats.oecd.org); and from the 2010 and 2011 Human Development Reports (United Nations Development Programme, 2010, 2011). Whenever possible, missing data for 2011 were replaced by the data from the preceding or the subsequent year (or the mean of both). For all measures, higher values indicate higher gender equity.

Analyses

Because of the multimatrix design, latent constructs were measured by plausible values. Generally speaking, plausible values result from applying missing value theory to estimates of latent construct values. The latent constructs in TIMSS and PIRLS are the reading, mathematics, and science competency values. These competency values are conceptualized as missing values. Therefore, multiple imputation strategies can be used to replace the missing values with reasonable estimates (Rubin, 1987). For this purpose, an imputation model including all variables that are part of the analysis model is needed to predict the most likely estimates.

For analyzing the relations across reading, mathematics, and science, a multidimensional IRT model was used as an imputation model (Martin & Mullis, 2012). As predictors, the students' scores on the test items as well as additional conditional variables were used. These conditional variables were principal components from a principal component analysis of all available student-level contextual data (e.g., family background, learning activities, domain-specific ability self-concept). To account for the uncertainty in the imputation strategy, five plausible values per student and domain were sampled and the variance of the statistic Q across the imputation values was added to the standard errors of Q (for a detailed description of the scaling procedure and the use of plausible values, see Martin & Mullis, 2012).[2] This scaling procedure preserved the correlational structure across the three subjects (multidimensional model). In addition, by using the conditional variables the measurement accuracy of the latent construct values increased, that is, the overall reliability of the three achievement scales was improved. The international median reliability was .82 for mathematics, .78 for science, and .88 for reading (Martin & Mullis, 2012). After executing the scaling procedure, each achievement scale was put on its own metric with an international mean of 500 and a standard deviation of 100.

To derive students' multidimensional proficiency profiles, we conducted a LPA (Gibson, 1966; Lazarsfeld & Henry, 1968). A

[1] Reading comprised reading intentions and reading comprehension. Mathematics comprised arithmetic, geometry, and handling data. Science comprised biology, physics/chemistry, and geography. Item examples can be obtained from http://timssandpirls.bc.edu/timss2011/international-released-items.html (TIMSS 2011) and http://timssandpirls.bc.edu/pirls2011/international-released-items.html (PIRLS 2011).

[2] Five plausible values were seen as optimal in terms of efficiency. More than five plausible values would have caused only marginally greater measurement accuracy (Schafer, 1999).

three-step approach was employed. In the first step, the total sample ($N = 74,868$) was separated into two to eight mixtures after we had weighted the data for the countries' different sample sizes so that every country contributed equally to the profile solution. The resulting seven mixture distributions were then compared based on the log-likelihood, the consistent Akaike information criterion, the Bayesian information criterion (BIC), and the sample-size-adjusted BIC (aBIC). After deciding for the mixture with the best data fit (Model 1; international model), the conditional means of the profiles were calculated and all students of the sample were assigned to the mixtures based on their most likely latent profile membership. Thus, the latent profiles were characterized by (a) their conditional means on the dimensions of the multidimensional distribution and (b) the percentage of students composing the profiles.[3]

In the second step, the conditional means were introduced as fixed parameters in national LPAs. That is, the country-specific multidimensional marginal distribution of students' achievement scores were separated into mixture distributions, in which both the number and the means of the mixtures were fixed at the values from the international model (Model 2; country-specific models with fixed means). Doing so allowed us to estimate the percentage of students forming the profiles and to assess how well the students from each national sample could be assigned to one of the profiles of the international model. The latter was indicated by both the relative entropy and the classification error rate.[4]

In the third step, Model 2 was applied again but, this time, the conditional means were free parameters; additionally, students' gender was included as a given manifest class (Model 3). Hence, the gender distribution within the national profiles could be examined. Due to the complex sampling procedure implemented in TIMSS and PIRLS, the students in the samples had had different probabilities of being selected. To adjust for these different selection probabilities and, thus, to make the sample representative of the desired population, we inversely weighted the units by their selection probabilities before applying Models 2 and 3.

Because five plausible values for the achievement domains were used, all analyses were performed five times (for each plausible value once). The results of these analyses were combined according to the formula by Rubin (1987). All analyses were conducted using Mplus 7.11 in combination with the full information maximum likelihood approach.

Results

Descriptive Statistics

Means and standard errors of the reading, mathematics, and science achievement test scores for the whole sample and for each country are displayed in Table 1 (see also Martin & Mullis, 2012; Mullis et al., 2012; Mullis, Martin, Foy, & Drucker, 2012). Although there was some variability across countries, girls consistently scored higher than boys in reading (average $d = 0.12$), and boys scored consistently but negligibly higher than girls in mathematics ($d = 0.06$). In science, there was also an average $d = 0.06$ in favor of boys, but the pattern of differences was rather inconsistent across countries.

Latent Profile Models

We first conducted an LPA using the samples of all 17 countries (weighted to be equal) to determine the international model (Model 1). Table 2 displays the fit criteria for the different possible international models. As can be seen, the consistent Akaike information criterion, the BIC, and the aBIC consistently decreased as the number of assumed profiles increased. Thus, the model with eight profiles achieved the best fit. However, as soon as the number of profiles exceeded seven, the number of students became extremely small for some profiles and therefore lacked substantive importance. In solutions with more than eight profiles, this pattern would have become even more extreme, given that the multidimensional distribution of the plausible values was approximately normal. Therefore, we chose the eight-profile solution as the upper bound of our analyses. When compared with the eight-profile model, the seven-profile model displayed somewhat more balanced proportions of students across the profiles, so that the numbers of students in the different profiles were of more practical importance (see Table 3, Columns 1 to 3). Because the difference in model fit between the eight-profile model and the seven-profile model was also very small and because the seven-profile model was the more parsimonious model and interpretable, we chose the seven-profile model as the international model.

In Hypothesis 1, we postulated that the international model would reveal student profiles that differ in their overall performance level across all three academic competencies rather than in the shape of their performance profile. In line with this prediction, the test scores of all achievement domains progressively and consistently increased from Profile 1 to Profile 7 (see Table 3, Columns 7, 9, and 11). No cross-nested structure for the profiles was observed. This suggests that the students could only be separated by different achievement levels on all three domains simultaneously and that no differentiation with respect to subject-specific strengths or weaknesses was possible.

In Hypothesis 2, we postulated that a reasonably large number of students in each country could reliably be assigned to one of the profiles of the international model built across all countries. We applied the international model in each country separately (with every country's data weighted to be representative), while holding the means constant (Model 2). The results indicated an acceptable assignment in 15 of the 17 countries (entropy $> .75$, classification error rate $< .25$). The assignment of the students from Romania and Malta was not as straightforward. More than a quarter of the students in these countries could not be assigned reasonably well into the international profile model. Nevertheless, a transfer from the international model to the country-specific distributions of students' achievement scores was possible without further restrictions. Thus, Hypothesis 2 was supported.

[3] We did not restrict the variances to be equal across the profiles.

[4] Relative entropy is the degree to which subjects can be clearly separated into the profiles. A value of 1 indicates a perfect fit, whereas a value of 0 indicates that the subjects could not be clearly separated into the profiles; classification error rate is the average posterior cross-classification probability. The posterior cross-classification probability of profile l is the likelihood that a student who is assigned to profile k will belong to profile l. Thus, the classification error rate represents the reliability of the profile solution. It ranges from 0 to 100% and is 0% when the profile solution is perfectly reliable.

Table 2
Comparison Between Models With Different Numbers of Latent Profiles (International Model; Model 1)

Number of profiles	Mean 2 × log-likelihood	CAIC	BIC	aBIC	Number of parameters
2	−1,266,986	2,533,992	2,534,084	2,534,052	10
3	−1,243,507	2,487,041	2,487,170	2,487,126	14
4	−1,231,360	2,462,756	2,462,756	2,462,865	18
5	−1,224,201	2,448,446	2,448,649	2,448,579	22
6	−1,220,405	2,440,862	2,441,102	2,441,020	26
7	−1,218,364	2,436,789	2,437,065	2,436,970	30
8	−1,217,217	2,434,503	2,434,816	2,434,708	34

Note. log-likelihood = the value of the corresponding fit-function when the estimation algorithm reaches the convergence criterion; CAIC = consistent Akaike information criterion; BIC = Bayesian information criterion; aBIC = sample-size-adjusted Bayesian information criterion. Smaller values indicate a better fit of the assumed mixture.

Gender Ratios in the Competency Profiles

To inspect the gender ratios (Research Question 3), we fixed the number of profiles and applied the model in every country, this time including gender in the analyses (Model 3). Overall, 73,331 of the 74,868 students (97.9%) could be assigned to one of the profiles. As can be seen in Table 3 (Column 6), boys were on average overrepresented at both the low and the high end and underrepresented in the middle range of the ability spectrum. Overall, boys outnumbered girls in Profile 1 by a ratio of 1.19 and in Profile 7 by a ratio of 1.24.

We then inspected variability in the gender ratios across the countries among the Profile 7 students (Research Question 4). Table 4 (Columns 2 to 4) shows that there was some considerable variation in the gender ratios. Although boys outnumbered girls in 14 of the 17 countries (even though the gender ratio of 1.15 in Lithuania was not statistically significant, $p > .01$), the gender ratios varied between 1.14 (Finland) and 1.67 (Czech Republic). Moreover, in two countries (Northern Ireland and Sweden), there was no difference between boys' and girls' percentage, and in one country (Ireland) there were slightly more girls than boys among the Profile 7 students (gender ratio: 0.96, *ns*).

To explain this variation (Hypothesis 5), we correlated the Profile 7 gender ratios with the values of the gender equity indicators (Table 4, Columns 5 to 8). The ratio of the net tertiary school enrollment rate (enrollment in postsecondary education programs with a duration of at least 2 or 3 years, qualifying for either direct labor force entry or academic high-skill professions) showed a small correlation with the gender ratios ($r = .19$).[5] This was however in the unexpected direction: In countries with a more equal tertiary enrollment rate, the gender ratios were more in favor of boys. The ratio of women to men with at least secondary education, women's labor market participation (regardless of hours worked) in the group of 15- to 64-year-olds, and women's share of research positions (head count, regardless of part- or full-time employment) correlated in the expected direction and in middle size with the gender ratios ($r = -.36$, $-.42$, and $-.33$, respectively). Thus, in countries where women are higher educated relative to men, show higher rates of labor market participation relative to men, and have a higher share of research positions, the gender ratios tended to be more balanced. Taken together in a multiple regression analysis, these indicators explained 28.7% of the gender ratio variance.

Discussion

In the present study, we investigated students' competency profiles and gender differences within these profiles. Furthermore, we inspected cross-national variability of the gender ratios among the top performers and tested whether this variability might be explained by societal gender equity. We achieved this by analyzing representative TIMSS and PIRLS 2011 data from 74,868 fourth graders from 17 European countries.

Interrelatedness of Academic Competencies

We derived an international profile model with seven profiles. As predicted in Hypothesis 1, the test scores in all achievement domains consistently increased from Profile 1 to Profile 7. Thus, the profiles unambiguously represented different ability levels simultaneously across different domains. This finding indicates that academic achievement in one domain is highly related to the achievement in the other two domains. This interrelatedness was so strong that possible relative domain-specific strengths and weaknesses as documented in other studies (e.g., Brunner et al., 2013; Wang et al., 2013) had no impact on the formation of the profiles when representative samples of students from a wide range of settings were investigated.

This gives a hint toward the importance of determinants shaping academic competencies across domains, such as the SES and the school environment. Professional intervention programs especially for low-SES children could be applied to foster their academic development. This could be accompanied by improvements of schools located in low-SES communities and of parent–school interaction. Chung (2015) and Kellaghan (2015) suggest a variety of opportunities to reach these goals, such as home support initiatives, early literacy and numeracy intervention programs, involving parents in school activities, or attracting highly qualified teachers to low-SES schools.

In line with Hypothesis 2, the profile patterns held across countries. Only in Malta and Romania were there slightly different profile patterns. Further analyses revealed that in these two countries, students with relative strengths and weaknesses appeared more often than in the other countries. Clarifying the reasons for the weaker interrelatedness of academic competencies in these countries might be an interesting purpose of future research.

Beyond the influence of the determinants already discussed, there might be further reasons for the interrelatedness we found. Different academic competencies might be needed when working on a specific competency test. For example, reading skills might be required for solving mathematics or science test items. Indeed, mathematics (and science) tasks often require students at least to some extent to read texts and to understand their meaning to grasp what is mathematically required (e.g., Abedi & Lord, 2001). Reading ability is also crucial in the process of gaining new knowledge in domains such as science or mathematics (Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999). In turn, knowledge acquired in different school domains might facilitate reading achievement because reading comprehension is derived from context information (Fukkink, 2005). This consideration has important

[5] Due to the small number of cases, we do not report *p* values but instead focus on the practical significance of the findings.

Table 3
Characteristics of the International Model (Model 1) Profiles: Number and Relative Frequency of Students (Overall and by Gender), Gender Ratios as Well as Means and Standard Errors of Students' Reading, Mathematics, and Science Scores

Profile	*n*	%			Gender ratio	Reading		Mathematics		Science	
		Overall	Girls	Boys		*M*	*SE*	*M*	*SE*	*M*	*SE*
1	783	1.0	1.0	1.1	1.19	251	27.5	273	31.3	243	26.2
2	2,993	4.0	3.7	4.4	1.13	345	23.5	358	20.8	343	23.7
3	8,514	11.4	11.2	11.7	1.01	419	16.9	419	13.6	417	15.8
4	18,193	24.3	24.6	23.7	.95	482	11.8	474	10.1	480	11.9
5	24,140	32.2	32.3	31.2	.97	538	9.0	527	9.1	538	9.2
6	16,681	22.3	22.3	22.5	1.03	592	7.3	580	7.8	594	7.4
7	3,564	4.8	4.9	5.4	1.24	655	7.0	644	8.2	657	7.2

Note. Gender ratio = number of boys for every girl; base-rate corrected average across countries after weighting to achieve nationally representative samples and after weighting countries' differences in sample size. Greater standard errors in Profiles 1 to 3 are primarily due to their smaller number of cases.

implications for fostering practices and instruction. If students are to be fostered in, for example, their mathematical ability, one should at the same time foster reading ability to achieve maximum learning results. It seems that high competency levels in several domains are a prerequisite for excellence in one particular domain. Therefore, instruction in a particular subject should not only focus on fostering the respective competency (e.g., fostering mathematical ability in mathematics instruction) but should be more holistic in the sense that different abilities should be focused on at the same time (e.g., fostering mathematical ability *and* reading ability in mathematics instruction).

One further question is whether the relations between the different domains might also be due to methodological factors, such as common method variance. However, if any, common method variance most often makes up an only negligible part of the overall variance (Jensen, 1998; Spector, 2006). This is the case even in questionnaire data which are actually thought of to be especially prone to factors causing common method variance such as social desirability, negative affectivity, or acquiescence (e.g., Spector, 2006). Thus, it seems unlikely that common method variance could explain a noticeable part of the interrelatedness between the different achievement domains that caused our LPA results.

Gender Differences in Academic Competencies

The differences in mean scores in all domains were small, supporting the gender similarities hypothesis (Hyde, 2005). However, we found that boys were overrepresented at both the low and the high end of the proficiency spectrum and underrepresented in the middle range. The more the competency profile departed from

Table 4
Gender Differences Among the Country-Specific Top Achieving Students (Profile 7; Model 3) as Well as Gender Equity Indicators

Country	Students in Profile 7		Gender ratio	Women's access to education		Ratio of women to men participating in the labor market	Women's share of research positions (%; head count)
	% of girls (*SE*)	% of boys (*SE*)		Ratio of women to men with secondary or higher education	Ratio of women to men enrolled in tertiary education		
Austria	5.3 (.5)	6.9 (.5)	1.30*	.78	1.18	.87	29.0
Czech Republic	2.4 (.3)	4.0 (.4)	1.67*	.98	1.42	.79	28.2
Finland	4.2 (.4)	4.8 (.4)	1.14*	1.00	1.23	.94	32.1
Germany	5.0 (.5)	6.5 (.6)	1.30*	.98	.93	.87	26.7
Hungary	4.3 (.4)	5.1 (.4)	1.19*	.96	1.32	.83	31.7
Ireland	5.5 (.5)	5.3 (.5)	.96	1.01	1.04	.81	32.4
Italy	2.9 (.4)	4.1 (.4)	1.41*	.86	1.43	.71	34.9
Lithuania	2.6 (.3)	3.0 (.3)	1.15	.96	1.47	.91	52.1
Malta	3.7 (.5)	5.0 (.5)	1.35*	.88	1.34	.53	26.9
Northern Ireland[a]	3.0 (.4)	3.0 (.4)	1.00	1.01	1.36	.85	37.7
Poland	2.9 (.3)	4.0 (.4)	1.38*	.95	1.55	.81	38.6
Portugal	3.8 (.4)	4.4 (.5)	1.16*	.96	1.19	.89	46.4
Romania	4.1 (.4)	4.7 (.4)	1.15*	.93	n/a	.78	46.1
Slovak Republic	4.4 (.4)	5.7 (.4)	1.30*	.93	1.55	.79	42.6
Slovenia	3.8 (.4)	4.8 (.4)	1.26*	.74	1.70	.90	36.4
Spain	4.0 (.4)	5.3 (.5)	1.33*	.94	1.23	.84	38.7
Sweden	5.2 (.5)	5.2 (.5)	1.00	1.01	1.51	.94	37.2

Note. Gender ratio = number of boys for every girl; n/a = not available.
[a] Gender equity data from the United Kingdom.
* Significantly ($p < .01$) different from 1.00.

the middle range, the more boys were represented. Thus, even when reading, mathematics, and science are considered simultaneously, boys are still overrepresented at the high end of the competency distribution. Besides gender differences in interests and career choices, this finding might contribute to an explanation for why there are more men than women among the top performers in societal key positions (Ceci et al., 2009).

We also found boys to be overrepresented among the weakest performing students. Against the background of boys being less successful than girls in the educational systems in many countries around the world (e.g., Spinath, Eckert, & Steinmayr, 2014; Voyer & Voyer, 2014), this is an important finding, too. Single studies have demonstrated that boys' lower school achievement is likely to be explained by personality and motivational factors (e.g., Duckworth & Seligman, 2006; Kessels & Steinmayr, 2013; Steinmayr & Spinath, 2008). As it is possible that factors at other levels also contribute to the underperformance of boys in school, further studies should investigate whether sociocultural factors explain the overrepresentation of boys at the low end of the ability distribution.

What causes the greater variability in competencies in boys compared to girls? One explanation might be that boys do not only differ more in competencies but already in other characteristics influencing those competencies. Whereas boys and girls do not systematically differ in their variance in different personality and motivational determinants of academic success (Steinmayr & Spinath, 2008), they differ in their variance in general mental ability (GMA). Studies using large representative samples have found that there are no or only negligible gender differences in mean GMA, but that males display greater variability than females, which then results in a male overrepresentation both in the upper and in the lower tail of the GMA distribution (e.g., Deary, Thorpe, Wilson, Starr, & Whalley, 2003; Johnson, Carothers, & Deary, 2008; Strand, Deary, & Smith, 2006). However, the gender ratio in the upper tail of the GMA distribution has been narrowed within the last decades (see Ceci et al., 2009; Lohman & Lakin, 2009). Thus, sociocultural factors, among others (Ceci et al., 2009; Hyde, 2014), are likely to contribute to the gender gap in the upper tail of the GMA distribution.

Likewise, we found clear hints that gender differences in the highest academic competency profile are to a substantial degree due to sociocultural factors. When we inspected the gender ratios among the top academic performers, one striking finding was that there was some considerable variation in these gender ratios across the countries. Moreover, we found evidence for a significant role of society's gender equity in both education and professional life. Gender ratios in favor of boys were smaller in countries where women had more secondary or higher educational levels, were more present in the labor market, and held more research positions. It would be interesting to investigate whether these sociocultural factors also contribute to gender differences in the upper tail of the GMA distribution.

The correlation between the tertiary school enrollment ratio and the gender ratios was however in an unexpected direction. This might be explained by a selection effect. In countries where female enrollment rates are lower than male enrollment rates, the girls and women who are enrolled are a more strictly ability-selected group than are boys or men in these countries. This could lead to more balanced gender ratios (when corrected for the base rate of females and males attending school). This interpretation is supported by studies showing that gender gaps especially among high mathematics and science achievers were more balanced in countries with higher power distance (i.e., with high segregation of social groups and a high tolerance toward inequity; see Reilly, 2012). Females in those countries might not only compose a relatively highly abled group but might also be more motivated than males to gain education to overcome their lower social status (Reilly, 2012).

Of course, correlations do not allow for causal conclusions. The causation might also work in the other direction than in the direction discussed above. For example, the fact that there are more women participating in the labor market and holding more research positions might also be a consequence—and not a cause—of smaller gender differences in mean or high levels of academic competencies (see, e.g., Ceci et al., 2009). However, as Else-Quest et al. (2010) already noted, the hypothesis of such causation cannot explain why gender differences in academic competencies occur in some countries but not in others. Other possibilities might be a reciprocal causation (see also Else-Quest et al., 2010) or the influence of a superordinate factor such as gender stereotypes (Miller, Eagly, & Linn, 2015). Answering the question of the causal direction would be an interesting and challenging issue for future research. In any case, the cross-national variability of the gender differences demonstrates that for the genesis of gender differences among the top performers, sociocultural factors definitely play an important role.

Limitations and Future Directions

In our study, we presented findings from large nationally representative samples of fourth graders from 17 European countries sharing the same supranational educational policy initiative. Despite the advantages of such a sample selection, it might be desirable for future studies to include as many culturally different countries as possible to evaluate the cross-national variability of the gender ratios at the high end of academic competencies even more comprehensively. Furthermore, societal gender equity could not explain the entire gender ratio variance. There must be additional factors at work causing the gender ratios. To unravel them, it could be useful to investigate more thoroughly which part of the gender ratios is due to differences in mean and which part is due to differences in variability, because the causes for differences in mean might be different from the causes for differences in variability (Humphreys, 1988).

As a final limitation, we studied achievement test scores only in reading, mathematics, and science. Although these three competencies are regarded as the three core academic competencies with the most predictive power of important life outcomes, it would have been desirable to include even more competencies taught at school.

Conclusions

We showed that (a) elementary school students across 17 European countries could (only) be clustered according to their achievement level across all three domains readings, mathematics, and science, and not according to their performance profile shape. We also showed that (b) boys were more likely than girls to perform at the top level on academic performance tests. However,

we found that (c) there was some cross-national variability in this tendency and that societal gender equity partly explained this variability. This speaks to an important role of sociocultural factors for the explanation of gender differences among the top academic performers.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 24,* 219–234. http://dx.doi.org/10.1207/S15324818AME1403_2

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D., & Tarelli, I. (2012). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland [Achievement profiles of fourth-graders in Germany]. In W. Bos, H. Wendt, O. Köller, & C. Selter (Eds.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 269–301). Münster, Germany: Waxmann.

Brunner, M.. Gogol, K. M., Sonnleitner, P., Keller, U., Krauss, S., & Preckel, F. (2013). Gender differences in the mean level, variability, and profile shape of student achievement: Results from 41 countries. *Intelligence, 41,* 378–395. http://dx.doi.org/10.1016/j.intell.2013.05.009

Brunner, M., Krauss, S., & Kunter, M. (2008). Gender differences in mathematics: Does the story need to be rewritten? *Intelligence, 36,* 403–421. http://dx.doi.org/10.1016/j.intell.2007.11.002

Byrnes, J. P., & Miller, D. C. (2007). The relative importance of predictors of math and science achievement: An opportunity-propensity analysis. *Contemporary Educational Psychology, 32,* 599–629. http://dx.doi.org/10.1016/j.cedpsych.2006.09.002

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin, 135,* 218–261. http://dx.doi.org/10.1037/a0014412

Chung, K. K. H. (2015). Socioeconomic status and academic achievement. In J. D. Wright (Series Ed.) & C. Byrne, P. Schmidt, & C. McBride-Chang (Vol. Eds.), *International encyclopedia of the social and behavioral sciences: Education* (pp. 924–930). Philadelphia, PA: Elsevier. http://dx.doi.org/10.1016/B978-0-08-097086-8.92141-X

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35,* 13–21. http://dx.doi.org/10.1016/j.intell.2006.02.001

Deary, I. J., Thorpe, G., Wilson, V., Starr, J. M., & Whalley, L. J. (2003). Population sex differences in IQ at age 11: The Scottish mental survey. *Intelligence, 31,* 533–542. http://dx.doi.org/10.1016/S0160-2896(03)00053-9

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98,* 198–208. http://dx.doi.org/10.1037/0022-0663.98.1.198

Else-Quest, N. M., & Grabe, S. (2012). The political is personal: Measurement and application of nation-level indicators of gender equity in psychological research. *Psychology of Women Quarterly, 36,* 131–144. http://dx.doi.org/10.1177/0361684312441592

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136,* 103–127. http://dx.doi.org/10.1037/a0018053

European Commission. (2015, August 24). Strategic framework: Education and training 2020. Retrieved from http://ec.europa.eu/education/policy/strategic-framework/index_en.htm

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62,* 61–84. http://dx.doi.org/10.3102/00346543062001061

Ferriman, K., Lubinski, D., & Benbow, C. P. (2009). Work preferences, life values, and personal views of top math/science graduate students and the profoundly gifted: Developmental changes and gender differences during emerging adulthood and parenthood. *Journal of Personality and Social Psychology, 97,* 517–532. http://dx.doi.org/10.1037/a0016030

Fukkink, R. G. (2005). Deriving word meaning from written context: A process analysis. *Learning and Instruction, 15,* 23–43. http://dx.doi.org/10.1016/j.learninstruc.2004.12.002

Gibson, W. A. (1966). Latent structure analysis and test theory. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 78–88). Chicago, IL: Science Research Associates.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Diversity: Culture, gender, and math. *Science, 320,* 1164–1165. http://dx.doi.org/10.1126/science.1154094

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research, 63,* 94–105. http://dx.doi.org/10.3102/00346543063001094

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45. http://dx.doi.org/10.1126/science.7604277

Helwig, R., Rozek-tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research, 93,* 113–125. http://dx.doi.org/10.1080/00220679909597635

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58,* 47–77. http://dx.doi.org/10.3102/00346543058001047

Humphreys, L. G. (1988). Sex differences in variability may be more important than sex differences in means. *Behavioral and Brain Sciences, 11,* 195–196. http://dx.doi.org/10.1017/S0140525X00049402

Hunt, J. (2016). Why do women leave science and engineering? *ILR Review, 69,* 199–226. http://dx.doi.org/10.1177/0019793915594597

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60,* 581–592. http://dx.doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65,* 373–398. http://dx.doi.org/10.1146/annurev-psych-010213-115057

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence. A new look at the old question. *Perspectives on Psychological Science, 3,* 518–531. http://dx.doi.org/10.1111/j.1745-6924.2008.00096.x

Kellaghan, T. (2015). Family and schooling. In J. D. Wright (Series Ed.) & C. Byrne, P. Schmidt, & C. McBride-Chang (Vol. Eds.), *International encyclopedia of the social and behavioral sciences: Education* (pp. 751–757). Philadelphia, PA: Elsevier. http://dx.doi.org/10.1016/B978-0-08-097086-8.92005-1

Kessels, U., & Steinmayr, R. (2013). Macho-man in school: Toward the role of gender role self-concepts and help seeking in school performance. *Learning and Individual Differences, 23,* 234–240. http://dx.doi.org/10.1016/j.lindif.2012.09.013

Kuncel, N. R., Ones, D. S., & Hezlett, S. A. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181. http://dx.doi.org/10.1037/0033-2909.127.1.162

Kurtz-Costes, B. (2015). Families as educational settings. In J. D. Wright (Series Ed.) & C. Byrne, P. Schmidt, & C. McBride-Chang (Vol. Eds.), *International encyclopedia of the social and behavioral sciences: Education* (pp. 731–737). Philadelphia, PA: Elsevier. http://dx.doi.org/10.1016/B978-0-08-097086-8.92004-X

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136,* 1123–1135. http://dx.doi.org/10.1037/a0021276

Lohman, D. F., & Lakin, J. M. (2009). Consistencies in sex differences on the Cognitive Abilities Test across countries, grades, test forms, and cohorts. *The British Journal of Educational Psychology, 79,* 389–407. http://dx.doi.org/10.1348/000709908X354609

Machin, S., & Pekkarinen, T. (2008). Assessment: Global sex differences in test score variability. *Science, 322,* 1331–1332. http://dx.doi.org/10.1126/science.1162573

Marsh, H. W., & Hau, K.-T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Educational Psychology, 96,* 56–67. http://dx.doi.org/10.1037/0022-0663.96.1.56

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Miller, D. I., Eagly, A. H., & Linn, M. C. (2015): Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology, 107,* 631–644. http://dx.doi.org/10.1037/edu0000005

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment framework.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39,* 21–43. http://dx.doi.org/10.1023/A:1018873615316

Organization for Economic Cooperation and Development (OECD). (2014). *PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science* (Rev. ed., Vol. 1). OECD Publishing, Paris.

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology, 114,* S138–S170. http://dx.doi.org/10.1086/589252

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PLoS ONE, 7,* e39904. http://dx.doi.org/10.1371/journal.pone.0039904

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology, 107,* 645–662. http://dx.doi.org/10.1037/edu0000012

Reynolds, M. R., Keith, T. Z., & Beretvas, S. N. (2010). Use of factor mixture modeling to capture Spearman's law of diminishing returns. *Intelligence, 38,* 231–241. http://dx.doi.org/10.1016/j.intell.2010.01.002

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138,* 353–387. http://dx.doi.org/10.1037/a0026838

Rindermann, H. (2007). The *g* factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-Tests across nations. *European Journal of Personality, 21,* 667–706. http://dx.doi.org/10.1002/per.634

Rost, D. H. (2013). *Handbuch Intelligenz* [Handbook of intelligence]. Weinheim, Germany: Beltz.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Hoboken, NJ: Wiley-Interscience. http://dx.doi.org/10.1002/9780470316696

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3–15. http://dx.doi.org/10.1191/096228099671525676

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75,* 417–453. http://dx.doi.org/10.3102/00346543075003417

Spector, P. E. (2006). Method variance in organizational research: Truth or urban legend? *Organizational Research Methods, 9,* 221–232. http://dx.doi.org/10.1177/1094428105284955

Spinath, B., Eckert, C., & Steinmayr, R. (2014). Gender differences in school success: What are the roles of students' intelligence, personality and motivation? *Educational Research, 56,* 230–243. http://dx.doi.org/10.1080/00131881.2014.898917

Steinmayr, R., Dinger, F. C., & Spinath, B. (2010). Parents' education and children's achievement: The role of personality. *European Journal of Personality, 24,* 535–550. http://dx.doi.org/10.1002/per.755

Steinmayr, R., Dinger, F. C., & Spinath, B. (2012). Motivation as a mediator of social disparities in academic achievement. *European Journal of Personality, 26,* 335–349. http://dx.doi.org/10.1002/per.842

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality, 22,* 185–209. http://dx.doi.org/10.1002/per.676

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences, 19,* 80–90. http://dx.doi.org/10.1016/j.lindif.2008.05.004

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive abilities test scores: A U.K. national picture. *The British Journal of Educational Psychology, 76,* 463–480. http://dx.doi.org/10.1348/000709905X50906

United Nations Development Programme. (2010). *Human development report 2010.* Houndmills, United Kingdom: Palgrave Macmillan.

United Nations Development Programme. (2011). *Human development report 2011.* Houndmills, United Kingdom: Palgrave Macmillan.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140,* 1174–1204. http://dx.doi.org/10.1037/a0036620

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101,* 817–835. http://dx.doi.org/10.1037/a0016127

Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science, 24,* 770–775. http://dx.doi.org/10.1177/0956797612458937

Received September 10, 2015
Revision received June 2, 2016
Accepted June 3, 2016 ■

Instructions to Authors
Journal of Educational Psychology
www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychologial Bulletin, 139,* 133–151. http://dx.doi.org/10.1037/a0028566

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace ineivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at http://www.apa.org/pubs/authors/instructions.aspx?item=6.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . .." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . .." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

Supplemental materials. APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the incoming editorial office at AConley@apa.org.

NEW RELEASES

from the American Psychological Association

A Practical Guide to Cultivating Therapeutic Presence

Shari M. Geller
2017. 248 pages. Hardcover.

List: $69.95 | APA Member/Affiliate: $49.95
ISBN 978-1-4338-2716-7 | Item # 4317441

Cognitive–Behavioral Therapy

SECOND EDITION
Michelle G. Craske
2017. 224 pages. Paperback.

Series: ***Theories of Psychotherapy Series***®
List: $24.95 | APA Member/Affiliate: $24.95
ISBN 978-1-4338-2748-8 | Item # 4317445

The Psychology of Juries

Edited by Margaret Bull Kovera
2017. 400 pages. Hardcover.

List: $69.95 | APA Member/Affiliate: $54.95
ISBN 978-1-4338-2704-4 | Item # 4318146

Activities for Teaching Statistics and Research Methods

A Guide for Psychology Instructors
Edited by Jeffrey R. Stowell and William E. Addison
2017. 192 pages. Paperback.

List: $39.95 | APA Member/Affiliate: $29.95
ISBN 978-1-4338-2714-3 | Item # 4316177

APA Handbook of Trauma Psychology

Volume 1. Foundations in Knowledge
Volume 2. Trauma Practice
Editor-in-Chief Steven N. Gold
2017. 1,168 pages. Hardcover.

Series: ***APA Handbooks in Psychology***®
List: $395.00 | APA Member/Affiliate: $195.00
ISBN 978-1-4338-2653-5 | Item # 4311531

Helping Couples on the Brink of Divorce

Discernment Counseling for Troubled Relationships
William J. Doherty and Steven M. Harris
2017. 400 pages. Hardcover.

List: $69.95 | APA Member/Affiliate: $54.95
ISBN 978-1-4338-2750-1 | Item # 4317446

Frailty, Suffering, and Vice

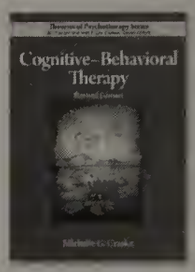
Flourishing in the Face of Human Limitations
Blaine J. Fowers, Frank C. Richardson, and Brent D. Slife
2017. 280 pages. Hardcover.

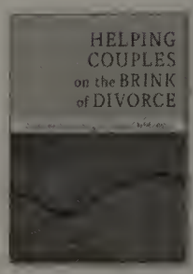
List: $69.95 | APA Member/Affiliate: $54.95
ISBN 978-1-4338-2753-2 | Item # 4317447

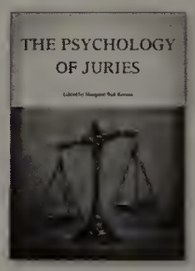
Mentalization-Based Treatment for Children

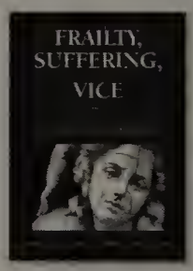
A Time-Limited Approach
Nick Midgley, Karin Ensink, Karin Lindqvist, Norka Malberg, and Nicole Muller
2017. 288 pages. Hardcover.

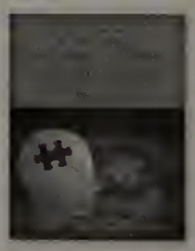
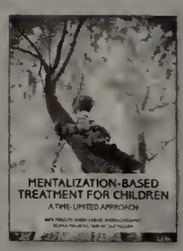
List: $69.95 | APA Member/Affiliate: $54.95
ISBN 978-1-4338-2732-7 | Item # 4317444







TO ORDER: 800-374-2721 • www.apa.org/pubs/books

AMERICAN PSYCHOLOGICAL ASSOCIATION | 1892 2017

AD3137